

Graph-Based Tests for Two-Sample Comparisons of Categorical Data

Hao Chen

Department of Statistics, Stanford University

Nancy R. Zhang

Department of Statistics, The Wharton School, University of Pennsylvania

Abstract:

We study the problem of two-sample comparison with categorical data when the contingency table is sparsely populated. In modern applications, the number of categories is often comparable to the sample size, causing existing methods to have low power. When the number of categories is large, there is often underlying structure on the sample space that can be exploited. We propose a general non-parametric approach that utilizes similarity information on the space of all categories in two sample tests. Our approach extends the graph-based tests of Friedman and Rafsky [1979] and Rosenbaum [2005], which are tests base on graphs connecting observations by similarity. Both tests require uniqueness of the underlying graph and cannot be directly applied on categorical data. We explored different ways to extend graph-based tests to the categorical setting and found two types of statistics that are both powerful and fast to compute. We showed that their permutation null distributions are asymptotically normal and that their p -value approximations under typical settings are quite accurate, facilitating the application of the new approach to real problems. The application of this new approach is illustrated through several examples.

Key words and phrases: Two-sample tests, categorical data, discrete data, minimum spanning trees, graph-based tests, contingency table.

1 Introduction

Testing whether two data samples are drawn from the same distribution is a fundamental problem in statistics. For low-dimensional Euclidean data, there are many approaches, both parametric and non-parametric, to this problem. When

the data are categorical, the existing approaches are much more limited. The standard procedure is to assume that each sample is drawn from a multinomial distribution, and the comparison becomes a test of whether the two samples come from the same multinomial distribution. Classical methods, such as the Pearson's Chi-square test and the deviance test, work well when we observe each category a large number of times. At least, the region in the contingency table where the two groups truly differ needs to be adequately sampled for existing tests to achieve good power. However, in many modern applications, the number of possible categories is comparable to or even larger than the sample size. Following are some examples:

Preference rankings: Survey data in marketing or psychometric research often come in the form of preference rankings. Subjects may be asked to rate wine (rank from best to worst tasting), pictures (choose 3 most familiar out of 5), or insurance plans (identify the most and least desirable). See Diaconis [1988] and Critchlow [1985] for more detailed examples on ranked and partially ranked data. It is a common problem to compare two groups of subjects to see if there is any between-group difference in preference. The number of possible full rankings is the factorial of the number of objects being rated, and the number of possible rankings is higher if some subjects only partially rank the objects.

Haplotype association: In genetics, a haplotype is a combination of alleles at adjacent loci on a chromosome that is transmitted together. A common problem of genetic association studies is to compare haplotype counts between treatment and control groups (e.g. see Zaykin et al. [2002] and Furihata et al. [2006]). Each haplotype can be represented as a fixed-length binary vector. The number of possible haplotypes is exponential in the number of loci. Haplotypes that are longer than 10 are often of interest in genetics, leading to $> 1,000$ possible combinations. However, the number of subjects in association studies is often only in the thousands or even hundreds, and the counts for most haplotypes are small.

Sequence or document comparisons: In the modern age of digitized texts, it is often of interest to compare the word composition in two different doc-

uments. A similar problem is the comparison of DNA or protein sequences, which plays a large role in bioinformatics [Lippert et al., 2002]. The number of possible words in these applications can be very large, while the counts for most words are small. For recent interest in this problem see Perry and Beiko [2010], Bush and Lahn [2006] and Rajan et al. [2007] for examples.

Classical Chi-square tests have low power in the above scenarios due to sparsity of the contingency table and high dimensionality of the parameter space. For exact tests, it is possible to generalize the concept to the setting of more than two categories, but this is computationally challenging [Mehta and Patel, 1983] and not efficient due to the existence in high dimensions of many equivalent tables, which are tables that have the same probability as the one observed.

When the number of categories is very large, there is often underlying similarity between different categories that can be exploited. For example, rankings can be related through Kendall’s or Spearman’s distance. Hamming distance or other more sophisticated measures can be used to compare haplotypes and fixed-length words in DNA sequences. In document comparison, the similarities between words are not equally likely: Some words are synonyms of others; Some are more likely to be used together. Such similarity information between categories can be properly used to improve the power of two-sample tests.

We assume that a distance matrix has been given on the set of categories, and adopt a graph-based approach proposed by Friedman and Rafsky [1979] and Rosenbaum [2005], where a graph is constructed on all subjects so that subjects more similar in value are connected by an edge. Friedman and Rafsky’s test is based on a minimum spanning tree (MST), and Rosenbaum’s test is based on minimum distance pairing (MDP). The test statistic in both cases is the number of edges connecting subjects from different groups. The underlying rationale is that, if two groups come from the same distribution, subjects coming from the same group should be as distant to each other as subjects coming from different groups. More details of these tests are given in Section 3. Both tests, however, require uniqueness of the underlying graphs. When the distance matrix on subjects is filled with ties, which is characteristic of categorical data, neither approach can be directly applied.

Ties in the distance matrix lead to ambiguity in constructing the MST or MDP, and the number of possible graphs increases rapidly with the number of ties. Some efforts were made to address this problem. In the analysis of a partially ranked data set with 38 subjects in 23 categories, Critchlow [1985] tried both the graph obtained from the union of all MSTs (uMST), and the graph obtained from the union of all nearest neighbor graphs (uNNG). Nettleton and Banerjee [2001] also used uNNG on a binary clinical feature data set with 64 subjects in 63 categories. In general, nearest neighbor graphs do not work well for categorical data, see Section 4. In this paper, Critchlow’s method using the uMST is studied in more detail and a computationally tractable form for categorical data is given. A different statistic, based on averaging over all optimal graphs of a certain kind, is also proposed and analyzed.

In Section 4, analytically tractable forms of the two statistics based on averaging over and union of minimum spanning trees are derived and compared via simulation to statistics based on MDP and uNNG. While the two MST-based tests are shown to be more powerful than the MDP- and uNNG-based tests, neither the averaged nor the union-based statistic dominate in power for the simulation scenarios explored. Algorithmic details for computing these two statistics are described, and in particular the averaged statistic is shown to be computationally intractable for some problems. A generalized version of the averaged statistic, with better computational properties, is proposed. In Section 5, the graph-based approach is illustrated on real and simulated data examples, and shown to have much better power than Chi-square tests. In Section 6, permutation null distributions of the proposed statistics are described. After mean- and variance- standardization, the statistics are shown to be asymptotically normal, under certain assumptions on the cell counts and the graph’s structure, as the number of observed categories goes to infinity.

2 Notations

We start by introducing our notations. The different categories are indexed by $1, 2, \dots, K$. The naming of the categories is arbitrary, that is, category 1 is not necessarily closer in distance to category 2 than to category 3. The two groups

are labeled a and b . The data is given in the form of a two-way contingency table (Table 1). Without loss of generality, we assume that each category has at least one subject over the two groups. That is, categories with no observation in either group can be omitted from the analysis without loss of information.

Table 1: Basic Notations.

	1	2	...	K	Total
Group a	n_{a1}	n_{a2}	...	n_{aK}	n_a
Group b	n_{b1}	n_{b2}	...	n_{bK}	n_b
Total	m_1	m_2	...	m_K	N

$$m_k = n_{ak} + n_{bk}, \quad k = 1, \dots, K;$$

$$n_a = \sum_{k=1}^K n_{ak}, \quad n_b = \sum_{k=1}^K n_{bk}, \quad N = n_a + n_b = \sum_{k=1}^K m_k.$$

Sometimes, we refer to individual subjects themselves, which we denote by Y_1, \dots, Y_N . Thus, each Y_i takes value in $\{1, \dots, K\}$ and has a group label

$$g_i = \begin{cases} a, & \text{if } Y_i \text{ belongs to group } a; \\ b, & \text{if } Y_i \text{ belongs to group } b. \end{cases} \quad (1)$$

We assume that a distance matrix, $\{d(i, j) : i, j = 1, \dots, K\}$ has been given on the set of possible categories, with $d(i, j)$ small if categories i and j are similar. Possible ways of defining the distance matrix are shown for various examples in Section 1.

A graph G is defined by its vertices and edges. We use G to refer to both the graph as well as its set of edges, when the vertex set is implicitly obvious. $|\cdot|$ is used to denote the size of the set, so $|G|$ is the number of edges in G . For any node i in the graph G , we use \mathcal{E}_i^G to denote the set of edges in G that contain node i , \mathcal{V}_i^G to denote the set of nodes in G that are connected to node i by an edge, and $\mathcal{E}_{i,2}^G$ to denote the set of edges in G that contain at least one node in \mathcal{V}_i^G . For any event A , I_A is the indicator function that takes value 1 if A is true and 0 otherwise.

Following is a list of abbreviations for different types of graphs and test statistics:

MST: Minimum Spanning Tree,

MDP: Minimum Distance Pairing,

NNG: Nearest Neighbor Graph,

uMST: The graph obtained by taking the union of all MSTs,

uNNG: The graph obtained by taking the union of all NNGs, equivalent to the graph connecting each point to all of its nearest neighbors,

R_G : The test statistic on the graph G ,

R_{aMST} : the test statistic averaging over all test statistics computed on each of the MSTs,

R_{aMDP} : the test statistic averaging over all test statistics computed on each of the MDPs,

3 A Review of Graph-Based Two-Sample Tests

By *graph-based two-sample tests*, we refer to tests that are based on graphs with the subjects $\{Y_i\}$ as nodes. We here suppose $\{Y_i\}$ take distinct values such that the graphs mentioned below can be constructed uniquely. The graph can be constructed using the distance matrix on $\{Y_i\}$. Friedman and Rafsky [1979] proposed the first graph-based two-sample test as a generalization of the Wald-Wolfowitz runs test to multivariate settings. Their test is based on a MST on the subjects, which is a spanning tree connecting all subjects that minimizes the sum of distances across edges. The Friedman-Rafsky test is based on the number of edges connecting subjects across different groups:

$$\sum_{(i,j) \in G} I_{g_i \neq g_j}, \quad (2)$$

where G is the MST. The statistic above is standardized to have mean zero and variance one, and its value is compared to the null distribution obtained by permuting the group labels. Friedman and Rafsky showed that, while this test has low power in low dimensions, it has comparable power to likelihood ratio

tests in a numerical study of normal data in > 20 dimensions, and higher power when the normal assumption is violated.

Another graph-based two-sample method, the cross-match test, was proposed by Rosenbaum [2005]. This test is based on minimum distance non-bipartite pairing (MDP), which divides the N subjects into $N/2$ (assuming N is even) non-overlapping pairs in such a way as to minimize the total of $N/2$ distances between pairs. For odd N Rosenbaum suggested creating a pseudo data point that has distance 0 with all other subjects, and later discarding the pair containing this pseudo point. The sum (2) is computed with G set to the MDP. The test statistic is the mean- and variance- standardized version of this sum. Note that the topology of the MDP does not depend on the distance matrix, with each node always having degree 1. This fact makes the test based on MDP truly distribution-free under the null hypothesis.

Both methods assume uniqueness of the type of graph used. For categorical data, ties appear in the distance matrix whenever a category has multiple counts. Even sparse contingency tables have quite a few cells containing more than one subject. The number of possible graphs grows rapidly with the number of ties. Thus, Friedman and Rafsky's and Rosenbaum's methods can not be directly applied to categorical data. For categorical data, distances are often based on qualitative measures, and thus while their relative ranking may be trustworthy, their absolute scale is not. Hence, we do not consider methods based directly on the distance matrix. While there are many ways to construct a graph based on a distance matrix, we limit our study to MST, MDP and uNNG, which are representative. Figure 1 illustrates the three different types of graphs on a simple example containing six points. These six points take on six distinct values.

4 Generalized Graph-Based Test Statistics

One natural solution, when the optimizing graph is not unique, is to average the test statistic over all graphs of the given kind. In this section, we consider the statistic based on averaging the sum (2) over all MSTs (R_{aMST}). Another solution to non-uniqueness is to take the union over all optimizing graphs, such as the statistic based on the uMST (R_{uMST}). R_{aMST} and R_{uMST} are analytically

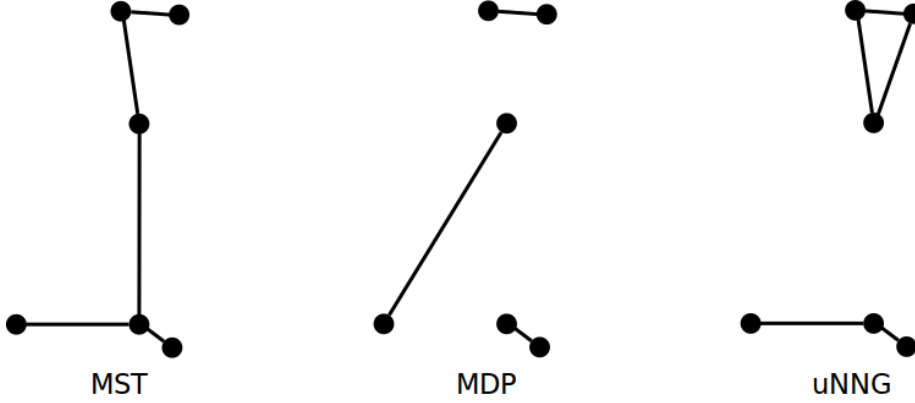


Figure 1: Illustration of MST, MDP, and uNNG on six points. Notice that there are two possible MSTs on the six points and only one is shown.

tractable and intuitively appealing, and their derivations are shown in Section 4.1. For comparison, we also consider the statistic based on averaging (2) over all MDPs, R_{aMDP} , and the statistic based on uNNG, R_{uNNG} . Computation of R_{aMDP} , described in Appendix A, is often intractable. Computation of uNNG is instantaneous. In Section 4.2, we study by simulation the performance of four graph-based statistics, R_{aMST} , R_{uMST} , R_{aMDP} , R_{uNNG} , comparing them to each other and to Chi-square tests. Our results show that tests based on minimum spanning trees have best power, and the intuition for this is explained. Computation of R_{aMST} and R_{uMST} is described in more detail in Section 4.3. When the number of MSTs on *categories* is large, which is common for categorical data, computation for R_{aMST} can be very costly. We generalize the statistic based on R_{aMST} to a similar but simpler form in Section 4.4.

4.1 The Test Statistics Based on MST

4.1.1 R_{aMST}

First, we define more notations. For each $k = 1, \dots, K$, let $\mathcal{C}_k \subset \{1, \dots, N\}$ be the subjects that belong to category k . From Table 1, $|\mathcal{C}_k| = m_k$. Let \mathcal{T}_k be the set of all spanning trees for \mathcal{C}_k . Since the distance between any two subjects in \mathcal{C}_k is zero, any spanning tree of \mathcal{C}_k is a MST of \mathcal{C}_k . Let \mathcal{T}_0^* be the set of all MSTs on the categories. We can embed each tree in \mathcal{T}_0^* as a graph on the subjects by

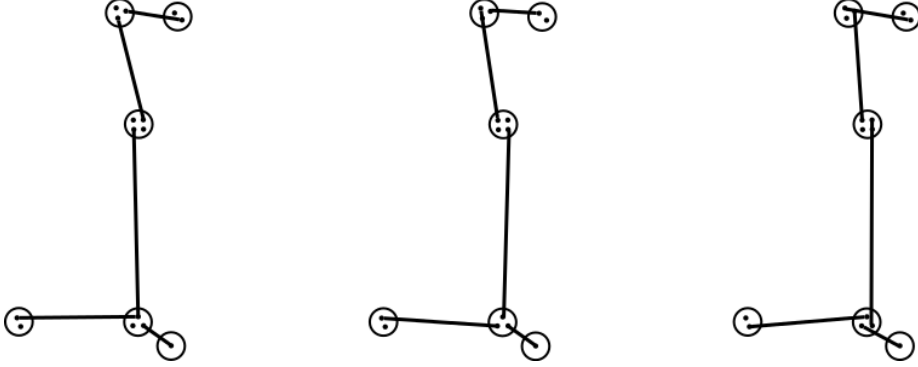


Figure 2: Embedding the MST on categories on the subjects. This figure only shows 3 out of 15552 possible embeddings.

randomly picking one subject in \mathcal{C}_k to represent category k , for $k = 1, \dots, K$. For each $\tau_0^* \in \mathcal{T}_0^*$, there are

$$\prod_{k=1}^K m_k^{|\mathcal{E}_k^{\tau_0^*}|} \quad (3)$$

different embeddings. For example, Figure 2 shows 3 out of 15552 ($= 2 \cdot 3^3 \cdot 1 \cdot 4^2 \cdot 3^2 \cdot 2$) possible embeddings for a MST on six categories containing 2, 3, 1, 4, 3 and 2 subjects. Let \mathcal{T}_0 be the set of all graphs obtained from embedding a tree from \mathcal{T}_0^* on the subjects. Then

$$|\mathcal{T}_0| = \sum_{\tau_0^* \in \mathcal{T}_0^*} \left(\prod_{k=1}^K m_k^{|\mathcal{E}_k^{\tau_0^*}|} \right). \quad (4)$$

Let \mathcal{T} be the set of all MSTs on the N subjects. Then, any member of \mathcal{T} can be represented as a union of a graph from \mathcal{T}_0 and a graph from each of $\{\mathcal{T}_k : k = 1, \dots, K\}$, and vice versa. Thus,

$$\mathcal{T} = \left\{ \tau_0 \cup \left(\bigcup_{k=1}^K \tau_k \right) : \tau_0 \in \mathcal{T}_0, \tau_k \in \mathcal{T}_k, k = 1, \dots, K \right\},$$

with

$$|\mathcal{T}| = |\mathcal{T}_0| \prod_{k=1}^K S_{m_k}, \quad (5)$$

where $S_m = m^{m-2}$ is the number of spanning trees on m points by Cayley's formula. Then, the test statistic based on averaging all MSTs on subjects can

be defined as:

$$R_{\text{aMST}} \triangleq |\mathcal{T}|^{-1} \sum_{\tau \in \mathcal{T}} R_{\tau}, \quad (6)$$

where R_{τ} is (2) with $G = \tau$. The following theorem gives a computationally tractable form for R_{aMST} in terms of the cell counts of the contingency table and the set of possible MSTs on categories.

Theorem 1. *The test statistic based on averaging over all MSTs on subjects is*

$$R_{\text{aMST}} = \sum_{k=1}^K \frac{2n_{ak}n_{bk}}{m_k} + |\mathcal{T}_0|^{-1} \sum_{\tau_0^* \in \mathcal{T}_0^*} \prod_{k=1}^K m_k^{|\mathcal{E}_k^{\tau_0^*}|} \sum_{(u,v) \in \tau_0^*} \frac{n_{au}n_{bv} + n_{av}n_{bu}}{m_u m_v}. \quad (7)$$

Proof.

$$\begin{aligned} R_{\text{aMST}} &= |\mathcal{T}|^{-1} \sum_{\tau \in \mathcal{T}} R_{\tau} \\ &= |\mathcal{T}|^{-1} \sum_{\tau_0 \in \mathcal{T}_0} \sum_{\tau_1 \in \mathcal{T}_1} \cdots \sum_{\tau_K \in \mathcal{T}_K} [R_{\tau_0} + R_{\tau_1} + \cdots + R_{\tau_K}] \\ &= |\mathcal{T}_0|^{-1} \sum_{\tau_0 \in \mathcal{T}_0} R_{\tau_0} + \sum_{k=1}^K \left[\sum_{\tau_k \in \mathcal{T}_k} R_{\tau_k} / S_{m_k} \right]. \end{aligned} \quad (8)$$

First consider the quantity $\sum_{\tau_k \in \mathcal{T}_k} R_{\tau_k} / S_{m_k}$. Since all pairs of subjects in a given category have the same distance ($= 0$), the edge between them should appear in the same number of trees. There are in total $m_k(m_k - 1)/2$ possible pairs and each spanning tree for \mathcal{C}_k has $m_k - 1$ edges. Hence, the edge between each pair of subjects in \mathcal{C}_k appears in exactly

$$\frac{S_{m_k}(m_k - 1)}{m_k(m_k - 1)/2} = \frac{2S_{m_k}}{m_k}$$

trees. Thus,

$$\sum_{\tau_k \in \mathcal{T}_k} \frac{R_{\tau_k}}{S_{m_k}} = \sum_{i,j \in \mathcal{C}_k: i < j} I_{g_i \neq g_j} \frac{2S_{m_k}/m_k}{S_{m_k}} = \frac{2n_{ak}n_{bk}}{m_k}. \quad (9)$$

Next consider the summation over \mathcal{T}_0 . For any $i \in \mathcal{C}_u$, $j \in \mathcal{C}_v$, if $(u, v) \in \tau_0^*$, then the edge (i, j) appears in

$$\prod_{k=1}^K m_k^{|\mathcal{E}_k^{\tau_0^*}|} / (m_u m_v)$$

elements in \mathcal{T}_0 , since any of the $m_u m_v$ possible edges connecting categories u and v appear in equal number of graphs in \mathcal{T}_0 . Thus,

$$\begin{aligned} \sum_{\tau_0 \in \mathcal{T}_0} R_{\tau_0} &= \sum_{\tau_0^* \in \mathcal{T}_0^*} \sum_{(u,v) \in \tau_0^*} \frac{\prod_{k=1}^K m_k^{|\mathcal{E}_k^{\tau_0^*}|}}{m_u m_v} \sum_{i \in \mathcal{C}_u} \sum_{j \in \mathcal{C}_v} I_{g_i \neq g_j} \\ &= \sum_{\tau_0^* \in \mathcal{T}_0^*} \prod_{k=1}^K m_k^{|\mathcal{E}_k^{\tau_0^*}|} \sum_{(u,v) \in \tau_0^*} \frac{n_{au} n_{bv} + n_{av} n_{bu}}{m_u m_v}. \end{aligned} \quad (10)$$

Combining (8), (9) and (10) gives (7). \square

The following corollaries show that R_{aMST} has a much simpler form if there is a unique MST on categories, or if the total number of subjects in each category is the same.

Corollary 1. *When $|\mathcal{T}_0^*| = 1$, then*

$$R_{\text{aMST}} = \sum_{k=1}^K \frac{2n_{ak}n_{bk}}{m_k} + \sum_{(u,v) \in \tau_0^*} \frac{n_{au}n_{bv} + n_{av}n_{bu}}{m_u m_v}, \quad (11)$$

where τ_0^* is the unique MST on categories.

Corollary 2. *When $m_k \equiv m$, $k = 1, \dots, K$,*

$$R_{\text{aMST}} = \sum_{k=1}^K \frac{2n_{ak}n_{bk}}{m} + |\mathcal{T}_0^*|^{-1} \sum_{\tau_0^* \in \mathcal{T}_0^*} \sum_{(u,v) \in \tau_0^*} \frac{n_{au}n_{bv} + n_{av}n_{bu}}{m^2}. \quad (12)$$

The form (11) of the statistic is especially intuitive. For each category k , we call the term $2n_{ak}n_{bk}/m_k$ the *mixing potential* of the category. The mixing potential is maximized if $n_{ak} = n_{bk} = m_k/2$, that is, when the subjects in category k are evenly divided between groups a and b ; it is minimized when the category contains subjects from only one group. A mixing potential for each edge (u, v) can also be defined as $(n_{au}n_{bv} + n_{av}n_{bu})/(m_u m_v)$. The edge-wise mixing potential is maximized when the edge connects a category containing only group a subjects with a category containing only group b subjects; it is minimized when both categories contain subjects only from one group. Thus, mixing potentials over categories and over edges between categories measure the similarity between the two groups. Corollary 1 shows that, when the MST on categories is unique, the test statistic R_{aMST} reduces to the sum of mixing potentials over nodes and

edges of the MST on categories. The similarity information on the categories is explicitly incorporated into the test through the sum of mixing potentials over the edges between categories.

In testing, the sums (7), (11) and (12) must be compared to their permutation distributions. A generalized statistic that we propose later in Section 4.4 is based directly on (11).

4.1.2 R_{uMST}

Following the notation from the previous section, let \mathcal{M}_0^* denote the set of edges appearing in at least one MST on categories. That is,

$$\mathcal{M}_0^* = \{(u, v) \in \tau_0^* : \tau_0^* \in \mathcal{T}_0^*\}.$$

In other words, \mathcal{M}_0^* is the uMST with the categories as nodes. When there is only one MST on categories, τ_0^* , then $\mathcal{M}_0^* = \tau_0^*$; when there are multiple MSTs on categories, which is common for categorical data, obtaining \mathcal{M}_0^* is not straightforward. Computation of \mathcal{M}_0^* is discussed in Section 4.3. The following theorem describes the analytic form of R_{uMST} given \mathcal{M}_0^* .

Theorem 2. *The test statistic based on uMST is*

$$R_{\text{uMST}} = \sum_{k=1}^K n_{ak}n_{bk} + \sum_{(u,v) \in \mathcal{M}_0^*} (n_{au}n_{bv} + n_{av}n_{bu}), \quad (13)$$

Proof. Within each category, every pair of subjects is connected, which gives the first term of (13). If categories u and v are connected in any $\tau_0^* \in \mathcal{T}_0^*$, then each point in category u is connected to every point in category v , giving the second term of (13). If categories u and v are not connected in any $\tau_0^* \in \mathcal{T}_0^*$, no edge will appear between categories u and v in uMST. \square

Remark 1. *Both R_{uMST} and R_{aMST} are derived from sums of $I_{g_i \neq g_j}$ over edges of the uMST on subjects. The main difference between them is that R_{uMST} treats all of the edges equally, while R_{aMST} assigns each edge a weight proportional to the number of MSTs on subjects in which the edge appears. Comparing (13) to (11), the denominators in (11) are omitted in (13). Each edge within category k appears in $|\mathcal{T}|/(m_k/2)$ MSTs, while each edge between categories appears in*

$|\mathcal{T}|/(m_u m_v)$ MSTs. Therefore, R_{uMST} puts more weight on between-category edges than within-category edges.

4.2 A Numerical Study

In this section, the power of the four tests based on R_{aMST} , R_{uMST} , R_{aMDP} and R_{uNNG} are studied and compared to Pearson’s Chi-square and deviance tests on simulated data sets. In each simulation, 30 points are randomly sampled from two different distributions – $N(0, 1)$ vs $N(1, 1)$, $N(0, 1)$ vs $N(0, 4)$, $N(0, 1)$ vs $N(1, 4)$, and $U(0, 5)$ vs $U(1, 6)$. The combined sample of 60 points is then discretized into 12 bins of equal width. The value 12 is chosen so that the average number of data points in each category is 5, mimicking the low cell count scenario. The bins are ranked by their start positions, and the distance between two categories is defined as the difference in their ranks. The p -values for all tests are calculated through 1,000 permutation samples for each simulation run, and the power is obtained from 1,000 simulation runs. In Figure 3, power is plotted versus type I error for each test and each simulation setting. Pearson’s Chi-square and deviance tests give very similar results, so only the results for the deviance test are shown. The deviance test is denoted by “LR” since it is based on the log-likelihood ratio. Power for all tests at the two most commonly used significance levels – 0.01 and 0.05 – are listed in Table 2.

First, compare R_{aMST} , R_{aMDP} , and R_{uNNG} . R_{aMST} is always significantly more powerful than R_{aMDP} , which in turn is always more powerful than R_{uNNG} . This result is intuitive from the definition of the different graphs. Since the MST must span the entire data set, $K - 1$ out of its $N - 1$ edges are forced to connect points between categories. For MDP, if a category has even number of subjects, the subjects in that category would be paired amongst themselves; between-category edges is only possible for those categories having an odd number of subjects. For uNNG, as long as a category has more than one subject, the subjects in that category would not be connected to subjects from other categories. Therefore, tests based on MST make the most use of the similarity information among categories, while the test based on R_{uNNG} makes the least use of this information. The simulation results show a positive correlation between using similarity information and the power of the test.

Now, we compare the test based on R_{uMST} to R_{aMST} . As discussed in Remark 1, the two statistics use the same set of edges but with different weighting. In simulation, the two statistics perform similarly under the three scenarios that compare two Normal distributions, while R_{uMST} has very little power, even much lower than R_{aMDP} and the deviance test, for the comparison of two Uniform distributions with different supports. When comparing two Normal distributions, the similarity between two categories is closely related to the difference of the ranks of the categories. That is, the further apart the ranks of the two categories, the less similar. However, when comparing two Uniform distributions with different supports – $[0,5]$ vs $[1,6]$ – only the ranks at the two ends are informative while the middle ranks are not. Since R_{uMST} puts more weight on between-category edges compared to R_{aMST} , it's power would be lower if the similarity measure among categories is not informative.

Note that of all the graph-based tests, only the test based on R_{aMST} consistently outperforms the deviance test.

N(0,1) vs N(1,1)	aMST	uMST	aMDP	uNNG	LR	Pearson
$\alpha = 0.01$	0.523	0.495	0.428	0.234	0.355	0.346
$\alpha = 0.05$	0.762	0.740	0.679	0.492	0.605	0.605
N(0,1) vs N(0,4)						
$\alpha = 0.01$	0.304	0.321	0.233	0.133	0.165	0.164
$\alpha = 0.05$	0.558	0.585	0.482	0.382	0.394	0.396
N(0,1) vs N(1,4)						
$\alpha = 0.01$	0.560	0.600	0.434	0.291	0.352	0.345
$\alpha = 0.05$	0.804	0.824	0.722	0.569	0.632	0.626
U(0,5) vs U(1,6)						
$\alpha = 0.01$	0.354	0.218	0.310	0.155	0.283	0.251
$\alpha = 0.05$	0.665	0.486	0.607	0.383	0.600	0.552

Table 2: The power of six tests – four graph-based tests based on R_{aMST} , R_{uMST} , R_{aMDP} , R_{uNNG} , the deviance test (LR) and Pearson's Chi-square test – under two significance levels ($\alpha = 0.01, 0.05$) and different simulation settings.

This simulation study is limited and only uses ranked data. We chose this study design for its interpretability. Though simple, the results are informative and show the advantage of averaged MST over averaged MDP and uNNG for

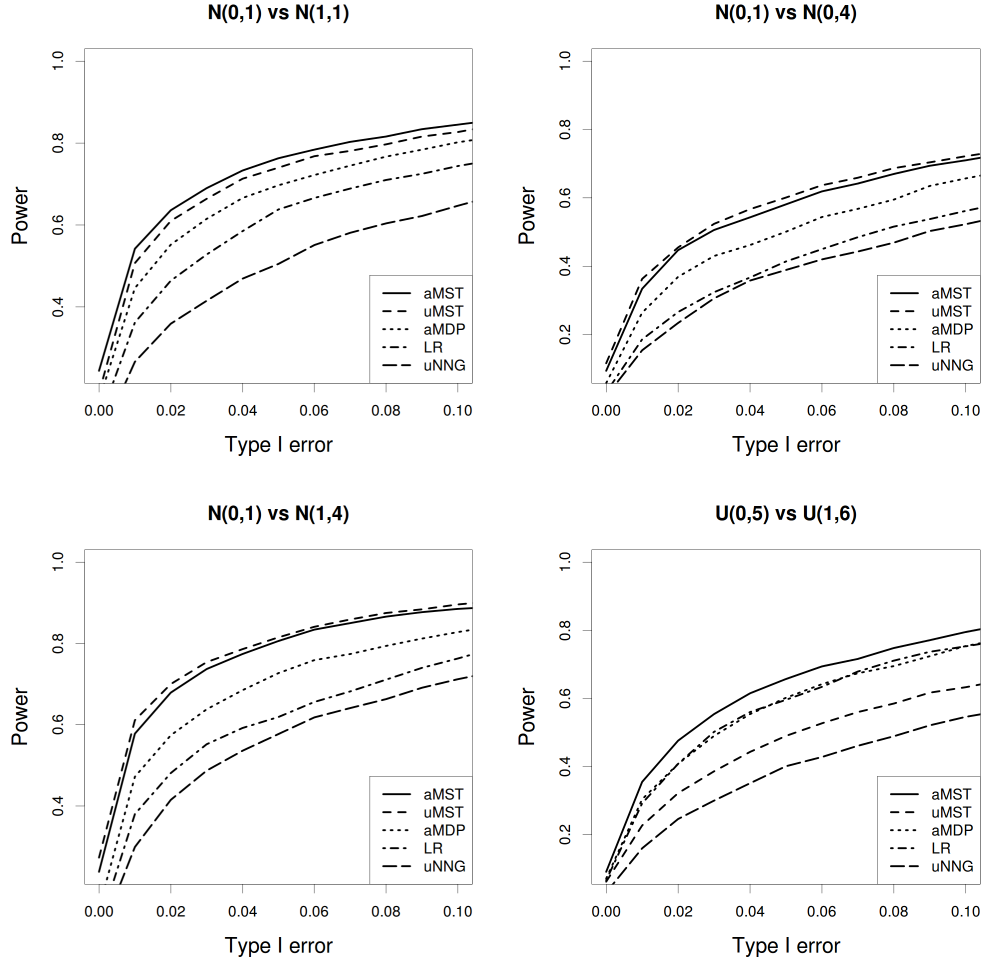


Figure 3: Power versus type I error for tests based on R_{aMST} , R_{uMST} , R_{aMDP} , the likelihood ratio (deviance), and R_{uNNG} under different simulation settings.

categorical data. Also, averaged MST is better than uMST when the similarity measure used to construct the graph is not effective. On the other hand, if the similarity measure is effective, the test based on uMST is comparable to, and sometimes better than, the test based on averaged MST. Hence, the rest of this paper focuses on the two tests based on R_{aMST} and R_{uMST} .

4.3 Computational Issues of R_{aMST} and R_{uMST}

The analytic forms for R_{aMST} and R_{uMST} , (7) and (13), require enumeration of all MSTs on categories for R_{aMST} ; and enumeration of all edges in \mathcal{M}_0^* for R_{uMST} . Let $M = |\mathcal{T}_0^*|$ be the number of MSTs on categories. If the distance matrix between categories is continuous-valued, then usually $M = 1$. Even when the distance matrix is arithmetic, M is small enough to be manageable for many problems. However, for problems that exhibit certain symmetries, enumeration of the set of all MSTs on categories is not computationally feasible. For example, Table 4.3 lists the values of M for the haplotype association problem in Section 5.2, assuming that there are no empty categories. In this problem, M is computed using the Matrix-Tree Theorem, yielding the formula

$$M = 2^{2^K - K - 1} \prod_{i=2}^K \exp \left\{ \binom{K}{i} \right\}.$$

For example, when the length of the haplotype is 6, which is a reasonably short length in genetic studies, there are only 64 possible categories, but M is already larger than 10^{45} . One may argue that in this case, (7) may be further simplified using the symmetry over categories, so that enumeration of $|\mathcal{T}_0^*|$ is not necessary. This is true if all categories are non-empty, but if one or more of the categories are empty, the symmetry breaks, while M would still be too large for enumeration.

Table 4 summarizes the computation time for R_{aMST} and R_{uMST} in terms of K and M . Consider first the listing of all edges in uMST on categories, \mathcal{M}_0^* , which is required for R_{uMST} . This task can be completed in $\mathcal{O}(K^2)$ time through an algorithm proposed by Eppstein [1995]. Details of the algorithm are in Appendix B, and its theoretical justification is completed by Chen [2012]. $\mathcal{O}(K^2)$ time is usually affordable since K is no larger than the sample size. Thus, R_{uMST} is computationally feasible for any problem. On the other hand, R_{aMST} requires the

Length of haplotype	K	M
2	4	4
3	8	384
4	16	42467328
5	32	2.078×10^{19}
6	64	1.66×10^{45}

Table 3: The number of categories, K , and the MSTs on categories, M , as haplotype length increases for the haplotype association problem in Section 5.2. All categories are assumed to be non-empty.

enumeration of all MSTs on categories, not just their edges, and thus adds $\mathcal{O}(M)$ computation time to the algorithm. For the haplotype example, this makes R_{aMST} computationally infeasible. Thus, in the next Section, we propose a statistic that is motivated by R_{aMST} but that is computationally tractable for all problems.

	Task	Computation Time
R_{aMST}	Enumerating all MSTs on categories	$\mathcal{O}(K^2 + M)$
R_{uMST}	Listing edges in uMST on categories	$\mathcal{O}(K^2)$

Table 4: Computational time for R_{aMST} and R_{uMST} . M is the number of MSTs on categories.

4.4 A Fast Method Generalized from R_{aMST}

Corollary 1 gives a simple and intuitive form of R_{aMST} when there is a unique MST on categories. In that special case, R_{aMST} is the sum of mixing potentials computed within each category and mixing potentials computed between categories that are connected by an edge of the MST τ_0^* . Evidence against the null increases if this sum of mixing potentials is small, as compared to random permutation. In (11), the MST τ_0^* serves as an enumeration of the pairs of categories that are highly similar. There is nothing sacred about the choice of MST for this role. The intuitive interpretation for (11) remains if we replace τ_0^* by any other graph C_0 that represents proximity between categories.

Up to this point, we have assumed that a distance matrix on categories is used to represent the similarity between categories. We now bypass the distance

matrix and assume that similarity is directly represented by a graph C_0 with the categories as nodes. Our goal is to incorporate the proximity information encoded by the graph into the two group comparison. We propose the following statistic, which we call R_{C_0} , obtained by substituting C_0 for τ_0^* in (11),

$$R_{C_0} = \sum_{k=1}^K \frac{2n_{ak}n_{bk}}{m_k} + \sum_{(u,v) \in C_0} \frac{n_{au}n_{bv} + n_{av}n_{bu}}{m_u m_v}. \quad (14)$$

The above statistic has a similar interpretation to R_{aMST} : Consider all C_0 -spanning graphs, which are graphs on subjects where every pair of subjects are connected by a path if they are in the same category or they are in two categories that are connected by a path in C_0 . Hence, minimum distance C_0 -spanning graphs connect subjects within categories by spanning trees and connects exactly one pair of subjects between each pair of categories that have an edge in C_0 . R_{C_0} is the averaged sum (2) over all minimum distance C_0 -spanning graphs.

If C_0 is given, computing R_{C_0} only requires $\mathcal{O}(K + |C_0|)$ time. If C_0 is not given, the choice of C_0 can often be guided by domain knowledge. In the examples below, our choices for C_0 include the uMST on categories, which we denote by C-uMST (same as \mathcal{M}_0^*), and the uNNG on categories, which we denote by C-uNNG. Since C-uMST and C-uNNG can both be computed in $\mathcal{O}(K^2)$ time, $R_{\text{C-uMST}}$ and $R_{\text{C-uNNG}}$ require only $\mathcal{O}(K^2)$ computation time for any problem.

5 Examples

In this section, the application of $R_{\text{C-uMST}}$, $R_{\text{C-uNNG}}$ and R_{uMST} are illustrated on several examples, both real and simulated. In the simulated examples, their power are compared to Chi-square tests. The p -values for all tests are calculated through 1,000 permutation samples for each run, and the power calculated through 1,000 simulation runs.

5.1 Preference Ranking

Consider comparing two groups of subjects on the ranking of four objects. Let Ξ be the set of all permutations of the set $\{1, 2, 3, 4\}$. Data are simulated under the following model: Subjects from group a have no preference among the four

objects, and so their ranking is uniformly drawn from Ξ . The rankings of subjects from group b are generated from the distribution

$$P_\theta(\zeta) = \frac{1}{\psi(\theta)} \exp\{-\theta d(\zeta, \zeta_0)\}, \quad \zeta, \zeta_0 \in \Xi, \theta \in \mathbb{R}, \quad (15)$$

where $d(\cdot, \cdot)$ is a distance function and ψ a normalizing constant. This probability model, first considered by Mallows [1957] with Kendall’s or Spearman’s distance, favors rankings that are similar to a modal ranking ζ_0 if $\theta > 0$. See Diaconis [1988] for more discussions. The larger the value of θ , the more clustering there should be in group b around the mode ζ_0 . We experimented with both Kendall’s and Spearman’s distance and various values for θ . We assumed that the true distance function used to generate the data is either known and used to construct the graph, or unknown, in which case an incorrect distance is used.

Figure 4 shows C-uMST and C-uNNG formed on a typical data set generated under $\theta = 5$ with $n_a = n_b = 20$. Spearman’s distance is used in both the generating model and for constructing the graph. In this particular example, C-uMST contains all edges in C-uNNG with three extra edges, shown in thinner lines. The reason this happens is that no category is as close to category “3241” as category “3142”, and no category is as close to category “3142” as category “3241”. For MST on categories, more edges are needed to form a spanning tree. It is clear that in this case, there are three MSTs on categories, each one obtained by adding one of the three thinner edges to the C-uNNG. In most simulation runs, C-uMST and C-uNNG are the same, while in those runs where they differ, C-uNNG is always a subset of C-uMST.

Figure 5 shows the power versus type I error for $\theta = 5, n_a = n_b = 20$ under different combinations of using Kendall’s or Spearman’s distance for the generating model and for constructing the graph; and Table 5 lists the power under two most commonly used significance levels – 0.01 and 0.05. We see that even when a wrong distance is used, the graph-based tests still have significantly higher power than the Chi-square tests. For this simulation setting, R_{uMST} is the most powerful among the three graph-based tests; $R_{\text{C-uMST}}$ and $R_{\text{C-uNNG}}$ perform similarly with $R_{\text{C-uMST}}$ a little better in all cases, implying that the extra edges in C-uMST do give additional useful information.

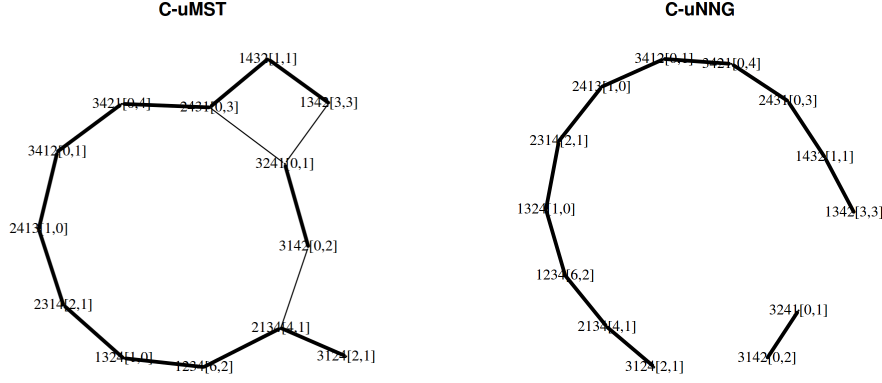


Figure 4: C-uMST and C-uNNG constructed on a typical data set generated under parameters $\zeta_0 = 1234$ and $\theta = 5$ with $n_a = n_b = 20$. The Spearman's distance is used in both the generating model and for constructing the graph. Each node is labeled by the ranking it represents, followed by the number of subjects from groups a and b with that ranking in parentheses.

KK	uMST	C-uMST	C-uNNG	Pearson	LR
$\alpha = 0.01$	0.566	0.413	0.397	0.214	0.206
$\alpha = 0.05$	0.784	0.660	0.648	0.450	0.439
KS					
$\alpha = 0.01$	0.567	0.395	0.385	0.221	0.209
$\alpha = 0.05$	0.784	0.649	0.631	0.455	0.437
SS					
$\alpha = 0.01$	0.588	0.491	0.478	0.247	0.240
$\alpha = 0.05$	0.807	0.715	0.703	0.485	0.480
SK					
$\alpha = 0.01$	0.607	0.495	0.486	0.253	0.248
$\alpha = 0.05$	0.811	0.729	0.715	0.494	0.481

Table 5: The power of five tests – three graph-based tests based on R_{uMST} , $R_{\text{C-uMST}}$, $R_{\text{C-uNNG}}$ and two Chi-square tests – under two significance levels ($\alpha = 0.01, 0.05$) and different simulation settings.

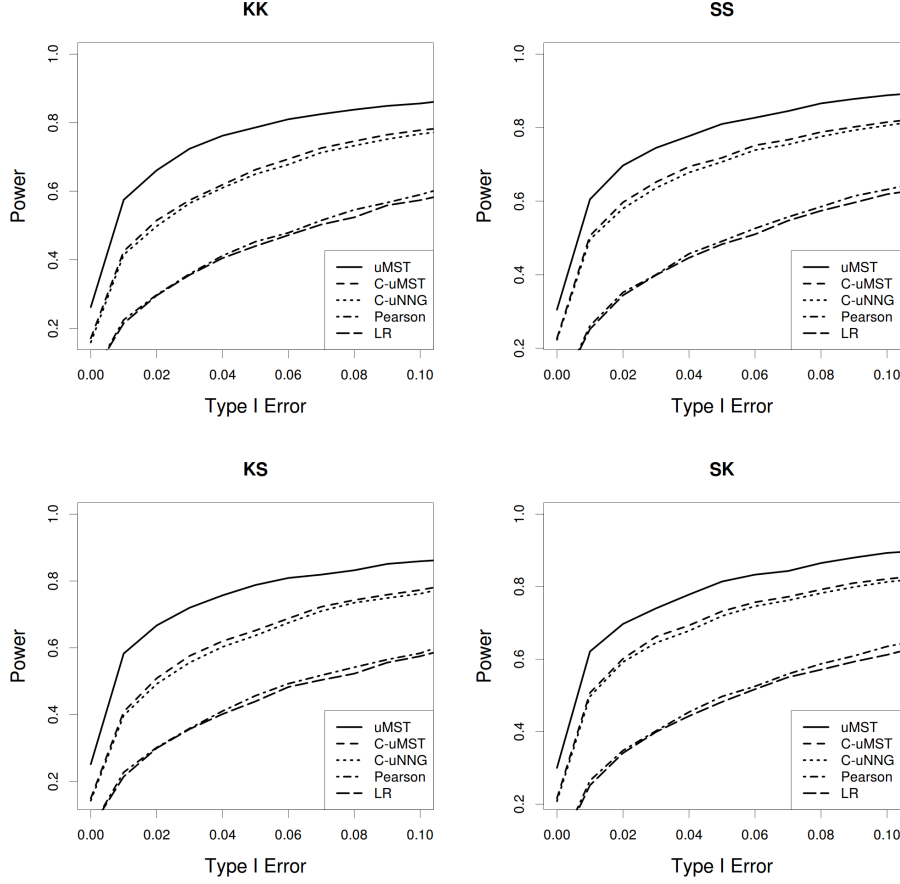


Figure 5: Power versus type I error for five different tests in the preference ranking example with $\theta = 5$ and $n_a = n_b = 20$. One of two distance measures (Kendall's or Spearman's distance) was used for the generating model and for constructing the graph. The title of each plot denotes the choice of distance: The first letter denotes the distance used in the generating model ("K" is Kendall's and "S" is Spearman's distance); and the second letter denotes the distance used in constructing the graph. For instance, "KS" in the bottom left panel means that Kendall's distance is used in the generating model, but Spearman's distance is used in constructing the graph.

5.2 Haplotype Association

In this example, we consider a disease model where the probability for disease depends on the haplotype at four single nucleotide polymorphisms (SNP). We encode the allele at each SNP as 0 or 1, and so the haplotype can be represented as a binary string. We assume that the disease probability depends on the number of positions at which the subject’s haplotype agrees with a target haplotype:

$$P(\text{Disease}) = 0.3 + 0.1 \times (\text{Number of positions in agreement}).$$

Thus, the probability of disease can take values 0.3, 0.4, 0.5, 0.6 or 0.7 depending on whether there are 0, 1, 2, 3 or 4 positions in agreement. To make the problem harder, we assume that seven non-informative SNPs are analyzed together with the four informative SNPs, and that which and how many of the 11 SNPs are informative is unknown in the analysis. Thus the data actually consists of haplotypes of length 11. There are $2^{11} = 2,048$ possible categories. In each simulation, 1,000 haplotypes with length 11 are generated uniformly from all possible haplotypes. Each subject with a given haplotype is signed as “patient” or “normal” according to the disease model. Since only 1,000 subjects are simulated in each run, not all of the 2,048 categories are represented. The number of non-empty categories in each run ranged from 755 to 823, with an average of 791 in the 1000 simulation runs. The Hamming distance is used to construct the graph. Figure 6 shows the power versus type I error plots for the five tests. It is clear that, by incorporating the information in the graph, tests based on R_{uMST} , $R_{\text{C-uMST}}$ and $R_{\text{C-uNNG}}$ all have much higher power than the Pearson’s Chi-square and deviance tests. Among the three graph-based tests, the one based on R_{uMST} works a little better than the ones based on $R_{\text{C-uMST}}$ and $R_{\text{C-uNNG}}$.

5.3 Binary Clinical Features

This example comes from Anderson et al. [1972] and Nettleton and Banerjee [2001]. Data on the presence or absence of 17 clinical features of the eye ailment Keratoconjunctivitis Sicca (KCS) are given for two groups of patients. A question asked by Nettleton and Banerjee was whether the two groups of patients share a common distribution with respect to these clinical features. The sizes of the groups are 40 and 24. It turned out that only two subjects had the same

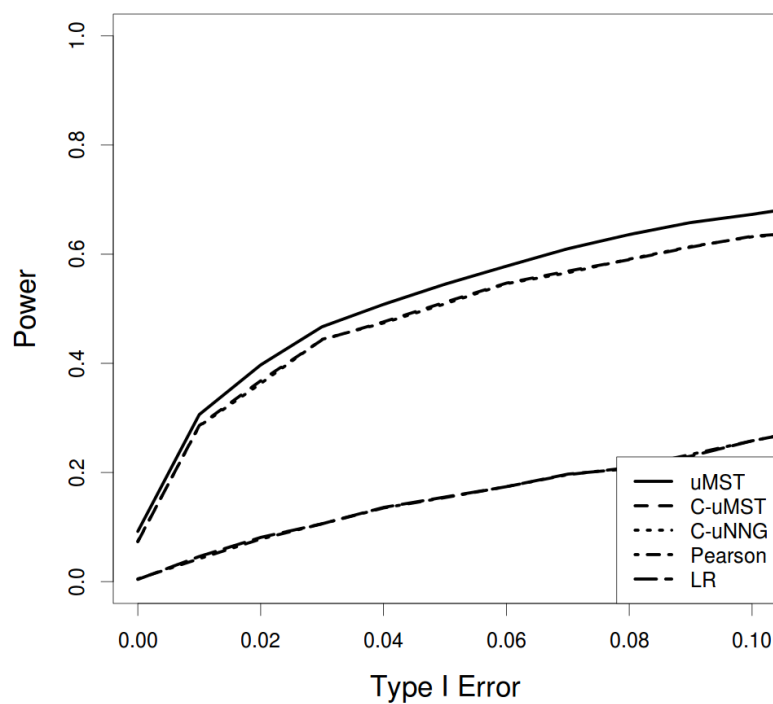


Figure 6: The power versus type I error plots for the five tests for the haplotype example. The length of the haplotype is 11, but only 4 positions informative.

outcome for the 17 clinical features, so there are in total 63 distinct categories. The Hamming distance is used to construct the graph, and p -values are calculated through 10,000 permutation samples and shown in Table 6. Nettleton and Banerjee's method is based on the uNNG on subjects. As discussed before and confirmed by simulation studies in Section 4.2, the uNNG on subjects has lower power than MST based tests when many categories have more than one subject. This is not a problem in this data because only one category has more than one subject. We see that R_{uMST} , $R_{\text{C-uMST}}$ and $R_{\text{C-uNNG}}$ all detected the difference between the two groups of patients, while the Chi-square tests did not.

R_{uMST}	$R_{\text{C-uMST}}$	$R_{\text{C-uNNG}}$	Nettleton and Banerjee's	Pearson	LR
0.0011	0.0010	0.0006	0.0007	0.5200	0.5200

Table 6: P -values for the KCS data set.

6 Permutation Distributions of the Test Statistics

Based on the results in Sections 4.2-4.4, we focus on $R_{\text{C-uMST}}$ and R_{uMST} . In this section, we consider the permutation distributions of these two statistics in their generalized forms. That is, we consider R_{C_0} and T_{C_0} , the latter defined as

$$T_{C_0} = \sum_{u=1}^K n_{au}n_{bu} + \sum_{(u,v) \in C_0} (n_{au}n_{bv} + n_{bu}n_{av}) \quad (16)$$

$T_{\text{C-uMST}}$ is equivalent to R_{uMST} .

We define two quantities that will be used to characterize the permutation distributions:

$$\lambda := \max_u |\mathcal{E}_u^{C_0}|, \text{ the maximum node degree in } C_0. \quad (17)$$

$$\beta := \max_u m_u, \text{ the maximum total count for a category.} \quad (18)$$

By permutation distribution, we are referring to the distribution of the statistic under random uniform permutation of the group labels. This is used as the null distribution to assess statistical significance. We use \mathbf{P}_p , \mathbf{E}_p and \mathbf{Var}_p to denote the probability, expectation and variance under the permutation null.

6.1 R_{C_0}

The following lemma states that the first two moments of R_{C_0} under the permutation null can be computed instantaneously using basic summary statistics of the graph and cell counts of the contingency table.

Lemma 1. *The mean and variance of R_{C_0} under the permutation null are*

$$\mathbf{E}_P[R_{C_0}] = (N - K + |C_0|)2p_1, \quad (19)$$

$$\begin{aligned} \mathbf{Var}_P[R_{C_0}] = & 4(p_1 - p_2) \left(N - K + 2|C_0| + \sum_{u=1}^K \frac{|\mathcal{E}_u|^2}{4m_u} - \sum_{u=1}^K \frac{|\mathcal{E}_u|}{m_u} \right) \\ & + (6p_2 - 4p_1) \left(K - \sum_{u=1}^K \frac{1}{m_u} \right) + p_2 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \\ & + (N - K + |C_0|)^2(p_2 - 4p_1^2), \end{aligned} \quad (20)$$

where

$$p_1 = \frac{n_a n_b}{N(N-1)}, \quad p_2 = \frac{4n_a(n_a-1)n_b(n_b-1)}{N(N-1)(N-2)(N-3)}. \quad (21)$$

Proof of Lemma 1 is given in Appendix C.1.

Remark 2. *As $N \rightarrow \infty$, $n_a/N \rightarrow \gamma \in (0, 1)$, we have $p_2 = 4p_1^2$ and thus*

$$\begin{aligned} \mathbf{Var}_P[R_{C_0}] = & 4(p_1 - p_2) \left(N - K + 2|C_0| + \sum_{u=1}^K \frac{|\mathcal{E}_u|^2}{4m_u} - \sum_{u=1}^K \frac{|\mathcal{E}_u|}{m_u} \right) \\ & + (6p_2 - 4p_1) \left(K - \sum_{u=1}^K \frac{1}{m_u} \right) + p_2 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}. \end{aligned}$$

Furthermore, if $\gamma = 0.5$, then $p_1 = p_2 = 1/4$ and we have

$$\mathbf{Var}_P[R_{C_0}] = \frac{1}{2} \left(K - \sum_{u=1}^K \frac{1}{m_u} \right) + \frac{1}{4} \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}.$$

We next give sufficient conditions guaranteeing the convergence to normality of R_{C_0} after standardization by its mean and variance.

Condition 1.

$$\sum_{u=1}^K m_u(m_u + |\mathcal{E}_u^{C_0}|)(m_u + \sum_{v \in \mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|) \sim o(K^{3/2}),$$

$$\sum_{(u,v) \in C_0} (m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|) (m_u + m_v + \sum_{w \in (\mathcal{V}_u \cup \mathcal{V}_v)} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|) \sim o(K^{3/2}).$$

Condition 1 constrains the size of “hubs” in the graph: The node degrees in C_0 and the number of observations in each category must not get too large. It can be simplified to stronger conditions that are easier to comprehend. For example, the following implies Condition 1:

Condition 1''. $\beta^6 \lambda^2$ and λ^8 are both $o(K)$.

The second condition is usually trivial:

Condition 2. N , $|C_0|$, and $\sum_{(u,v) \in C_0} \frac{1}{m_u m_v}$ are all $\mathcal{O}(K)$.

The asymptotic distribution of the standardized form of R_{C_0} is given in the following theorem.

Theorem 3. *Assume that conditions 1 and 2 hold. Under the permutation null, the standardized statistic*

$$\frac{R_{C_0} - \mathbf{E}_P[R_{C_0}]}{\sqrt{\mathbf{Var}_P[R_{C_0}]}}$$

converges in distribution to $N(0, 1)$ as $K \rightarrow \infty$ and n_a/N is bounded away from 0 and 1.

The proof of Theorem 3 is given in Appendix C.2.

Theorem 3 can be applied to any type of graph, allowing for repeated observations of each node. Since the statistics in Friedman and Rafsky [1979] and Rosenbaum [2005] do not allow ties, their asymptotic normality results are also restricted to the case where each node is observed only once. To compare Theorem 3 to its counterpart in these two papers, we let $G = C_0$ and assume that $m_u \equiv 1$. Thus,

$$N = K, \quad \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} = |C_0| = |G|.$$

Condition 2 requires that $|G| \sim \mathcal{O}(K)$ and Condition 1 can be simplified to:

$$\sum_{u=1}^K |\mathcal{E}_u^G| |\mathcal{E}_{u,2}^G| \sim o(K^{3/2}),$$

$$\sum_{(u,v) \in G} (|\mathcal{E}_u^G| + |\mathcal{E}_v^G|) (|\mathcal{E}_{u,2}^G| + |\mathcal{E}_{v,2}^G|) \sim o(K^{3/2}).$$

Hence, Theorem 3 implies the asymptotic normality result in Rosenbaum [2005] since $|\mathcal{E}_u^G| \equiv 1, |\mathcal{E}_{u,2}^G| \equiv 1, |G| = K/2$ for MDP. Friedman and Rafsky proved a more general condition for asymptotic normality of sums (2) after standardization: For sparse graphs where $|G| \sim \mathcal{O}(K)$, the number of edge pairs that share a common node must be $\mathcal{O}(K)$. Condition 1 is neither stronger or weaker than Friedman and Rafsky's condition. For example, consider a graph with one node having degree $K^{1/2}$ and all other nodes having degree 1; this graph satisfies Friedman and Rafsky's condition but not Condition 1, since $\sum_{(u,v) \in G} |\mathcal{E}_u^G| |\mathcal{E}_v^G| = \mathcal{O}(K^{3/2})$. On the other hand, a graph with \sqrt{K} nodes having degree $K^{0.3}$ and all other nodes having degree 1 would satisfy Condition 1 but not Friedman and Rafsky's condition.

6.2 T_{C_0}

The following lemma is the counterpart of Lemma 1 for R_{C_0} . It's proof is given in Appendix C.3.

Lemma 2. *The mean and variance of T_{C_0} under the permutation null are*

$$\mathbf{E}_P[T_{C_0}] = \left(\sum_{u=1}^K m_u(m_u - 1) + 2 \sum_{(u,v) \in C_0} m_u m_v \right) p_1, \quad (22)$$

$$\begin{aligned} \mathbf{Var}_P[T_{C_0}] &= (p_1 - p_2) \sum_{u=1}^K m_u(m_u + \sum_{v \in \mathcal{V}_u} m_v - 1)(m_u + \sum_{v \in \mathcal{V}_u} m_v - 2) \\ &\quad + (p_1 - p_2/2) \left(\sum_{u=1}^K m_u(m_u - 1) + 2 \sum_{(u,v) \in C_0} m_u m_v \right) \\ &\quad + (p_2 - 4p_1^2) \left(\sum_{u=1}^K m_u(m_u - 1) + 2 \sum_{(u,v) \in C_0} m_u m_v \right)^2, \end{aligned} \quad (23)$$

where p_1 and p_2 are given in (21).

The next theorem gives a sufficient condition for asymptotic normality of T_{C_0} under the permutation null.

Theorem 4. *If $\sum_{u=1}^K m_u(m_u + \sum_{v \in \mathcal{V}_u} m_v)^2 \sim \mathcal{O}(N)$, then under the permutation null distribution, the standardized statistic*

$$\frac{T_{C_0} - \mathbf{E}_P[T_{C_0}]}{\sqrt{\mathbf{Var}_P[T_{C_0}]}} ,$$

where $\mathbf{E}_P[T_{C_0}]$ and $\mathbf{Var}_P[T_{C_0}]$ are given in (22) and (23), converges in distribution to $N(0, 1)$ as $N \rightarrow \infty$, and n_a/N bounded away from 0 and 1.

Proof. Let \bar{G} be the uMST on subjects. Then as long as $\sum_{i=1}^N |\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}| - 1) \sim \mathcal{O}(N)$, asymptotic normality can be ensured following Friedman and Rafsky [1979]’s result. Notice that if i is in category u , then $|\mathcal{E}_i^{\bar{G}}| = (m_u - 1) + \sum_{v \in \mathcal{V}_u} m_v$. \square

6.3 Checking the P -values Under Normal Approximations

We now check the normal approximations to the p -values of the three graph-based statistics – $R_{C\text{-uMST}}$, $R_{C\text{-uNNG}}$ and R_{uMST} – through simulation. We adopt the setting of the haplotype example. In each simulation run, N haplotypes with length l are generated uniformly from all possible haplotypes with length l . They are assigned to either group with equal probability. Hence, the two groups have the same distribution. For each simulation run, we calculate the difference between theoretical p -values from the normal approximation and the permutation p -values from 10,000 permutations for the three statistics. We consider different sparsity settings by varying l , which controls the number of categories, and N . Under each setting, 100 simulation runs are done, with boxplots of the differences between theoretical and simulation p -value shown in Figure 7. We increased l from 6 to 10, and thus the number of possible categories considered grows from 64 to 1024. The sample size N varies from 100 to 1000. This spectrum of values is reasonable for a genetic association study.

Simulation results under this setting shows that the normal approximation is better for $R_{C\text{-uMST}}$ and $R_{C\text{-uNNG}}$ than for R_{uMST} . Accuracy of normal approximation improves for all statistics as l and N increase. For $R_{C\text{-uMST}}$ and $R_{C\text{-uNNG}}$, when the number of possible categories is larger than 256 and the number of observations is larger than 200, the p -value from normal approximation is quite accurate. While for R_{uMST} , the number of observations needs to be larger than 500 to

achieve similar accuracy. For R_{uMST} , when the number of possible categories is larger than the number of observations, the p -value calculated from the normal approximation is negatively biased, and thus less conservative. The bias is less severe for $R_{\text{C-uMST}}$ and $R_{\text{C-uNNG}}$, while still problematic when the number of possible categories is 1024 and the number of observation is only 100. Skewness correction can be done to make the theoretical p -values more accurate, but when N is so small, it would be easier to just do permutation directly.

7 Conclusions and Discussion

We have described a new approach for comparing two categorical samples, which is appealing when the contingency table is sparsely populated. Sparse contingency tables are common in many modern applications where the number of categories, K , is large compared to sample size. In such situations, the different categories can usually be related to each other in a systematic way. The new approach utilizes a graphical encoding of the similarity between categories to improve the power of two-sample comparison. We showed, through simulations and real examples, that utilizing graphical information improves the power over deviance and Pearson’s Chi-square tests. The proposed statistics are shown to be asymptotically normal after standardization, under assumptions that limit the hub size and density of the graph. This allows instantaneous type I error control for large data sets.

The power of the new approach depends on the choice of an informative similarity measure between categories. This part of the analysis should rely on domain knowledge that is specific to each application. For ranking data from surveys, one can start with the standard distance measures used in Example 5.1. When the number of categories is large, drawing relationships between categories is a necessary and often default step in analyzing the data.

Both $R_{\text{C-uMST}}$ and R_{uMST} work well when the similarity information is effective with R_{uMST} usually having better power. However, when the similarity measure is not as informative, R_{uMST} can have very low power, even when compared to Chi-square tests. In our simulation studies derived from the Haplotype problem, the normal approximation is more accurate for $R_{\text{C-uMST}}$ than for R_{uMST} . For R_{uMST} ,

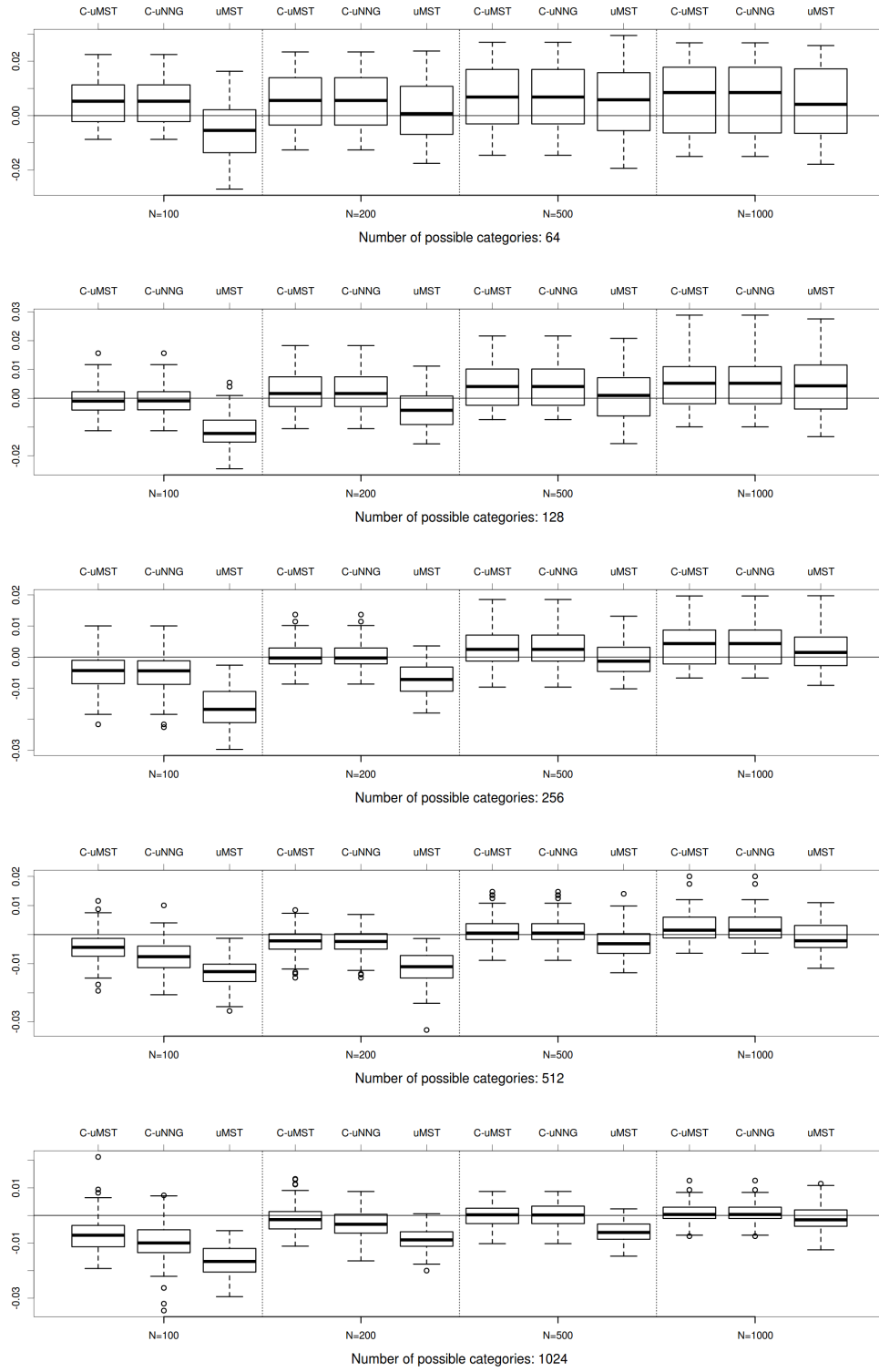


Figure 7: Boxplots for the differences between p -values calculated from normal approximation and 10,000 permutations.

p -values obtained from normal approximation are lower than actual p -values in extremely sparse situations. All p -value approximations work well when the sample size is comparable to the number of categories.

Generalization of this approach to multi-sample comparison is straightforward by letting g_i take K' distinct values, where K' is the number of groups.

A The Test Statistic Based on R_{aMDP}

We first assume N , the total number of observations, to be even. Let K_0 be the number of categories containing an odd number of subjects. Since N is even, K_0 is even. (K_0 can be 0.). Without loss of generality, let categories $1, \dots, K_0$ be the categories containing an odd number of subjects, and categories $K_0 + 1, \dots, K$ be the categories containing an even number of subjects. More notations are defined below.

- $\mathcal{A} = \{\mathbf{x} = (x_1, \dots, x_{K_0})^T : x_i \in \{a, b\}, i = 1, \dots, K_0\}$: all possible combinations of group identities of the subjects with one from each of the categories containing an odd number of subjects.
- $R_0(n_a, n_b)$: the number of edges connecting subjects from different groups averaged over all perfect pairings of n_a points from group a and n_b points from group b in the same category, with $n_a + n_b$ being even.
- $R_{\mathbf{x}}, \mathbf{x} \in \mathcal{A}$: the number of edges connecting subjects from different groups averaged over all MDPs on categories $1, \dots, K_0$.

Assumption 1. *If a category has an even number of subjects, the subjects are paired within the category.*

Assumption 1 is usually true for MDP on subjects for categorical data. It is explicitly stated here to avoid the complicated scenario when the triangle inequality becomes equality in the distance metric for any three categories.

Proposition 1. *Under Assumption 1, the test statistic based on averaging (2)*

over all MDPs is:

$$R_{aMDP} = \sum_{k=K_0+1}^K R_0(n_{ak}, n_{bk}) + \frac{1}{\prod_{k=1}^{K_0} m_k} \sum_{\mathbf{x} \in \mathcal{A}} \left\{ \prod_{i=1}^{K_0} n_{x_i i} \left[R_{\mathbf{x}} + \sum_{j=1}^{K_0} R_0(n_{x_j j} - 1, n_{x_j^c j}) \right] \right\}, \quad (24)$$

$$\text{where } x_i^c = \begin{cases} b & \text{if } x_i = a \\ a & \text{if } x_i = b \end{cases},$$

$$R_0(n_a, n_b) = \sum_{i \in \mathcal{S}} i \binom{n_a}{i} \binom{n_b}{i} i! (n_a - i - 1)!! (n_b - i - 1)!! / (n_a + n_b - 1)!! \quad (25)$$

with

$$\mathcal{S} = \begin{cases} \{0, 2, \dots, n_a \wedge n_b\} & \text{if } n_a \text{ and } n_b \text{ both even} \\ \{1, 3, \dots, n_a \wedge n_b\} & \text{if } n_a \text{ and } n_b \text{ both odd} \end{cases},$$

and

$$R_{\mathbf{x}} = |\Omega^*|^{-1} \sum_{\omega^* \in \Omega^*} \sum_{(i,j) \in \omega^*} I_{x_i \neq x_j}, \quad (26)$$

where ω^* is an MDP on categories $1, \dots, K_0$, and Ω^* is the set of all these ω^* 's.

Proof. First consider the simpler case: One category with n_a subjects from group a and n_b subjects from group b , with $n_a + n_b$ even. Since all subjects are in the same category, any perfect pairing is an MDP. There are in total $(n_a + n_b - 1)!!$ different perfect pairings.

When both n_a and n_b are even, the possible numbers of edges connecting different groups are: $0, 2, \dots, n_a \wedge n_b$. Among all the $(n_a + n_b - 1)!!$ perfect pairings, the number of perfect pairings having $i \in \{0, 2, \dots, n_a \wedge n_b\}$ edges connecting different groups is

$$\binom{n_a}{i} \binom{n_b}{i} i! (n_a - i - 1)!! (n_b - i - 1)!!.$$

When both n_a and n_b are odd, the possible numbers of edges connecting different groups are: $1, 3, \dots, n_a \wedge n_b$. Among all the $(n_a + n_b - 1)!!$ perfect pairings, the

number of perfect pairings having $i \in \{1, 3, \dots, n_a \wedge n_b\}$ edges connecting different groups is also

$$\binom{n_a}{i} \binom{n_b}{i} i! (n_a - i - 1)!! (n_b - i - 1)!!.$$

(25) follows immediately.

Under Assumption 1, an MDP on all subjects would be an MDP on categories $1, \dots, K_0, (\omega^*)$, embedded on the subjects similar to the MST case, with all other subjects paired within each category, so (24) follows naturally. \square

Remark 3. *If N , the total number of observations, is odd, we can add a pseudo category with one subject, whose distance to any other category is 0. All derivations are the same, except that the edge containing the pseudo category is discarded from the MDP on categories in later steps.*

B Computation Time for R_{aMST} and R_{uMST}

The main task for computing R_{aMST} and R_{uMST} are to enumerate all MSTs on categories for R_{aMST} and to list the edges in \mathcal{M}_0^* for R_{uMST} . Other tasks can be finished in $\mathcal{O}(K)$ time.

Let G be the complete graph on the K categories. $|G| = K(K - 1)/2$. Eppstein [1995] proposed a graph operation called the sliding transformation which, when applied to G , produces an equivalent graph such that the MSTs on categories correspond one-for-one with the spanning trees of the equivalent graph. The enumeration of all spanning trees, without having to optimize for total distance, is relatively straightforward. Thus, we adopted the following computational approach: Use Eppstein's method to construct the equivalent graph of G , enumerate all spanning trees of the equivalent graph, then transform back to get the set of MSTs on G . The sliding transformation constructs the equivalent graph in $\mathcal{O}(|G| + K \log K) = \mathcal{O}(K^2)$ time. To perform the sliding transformation, an initial MST is needed. Prim's algorithm can be used to obtain the initial MST, which needs $\mathcal{O}(K^2)$ time, not increasing the time complexity. The theoretical justification of this algorithm can be found in Eppstein [1995] and Chen [2012], which completes many of the proofs of Eppstein [1995].

After removing any loops formed during the the sliding transformations, each remaining edge appears in at least one spanning tree of the equivalent graph, thus appearing in at least one MST on G . Now we have the list of edges in uMST on G , and thus R_{uMST} can be calculated in $\mathcal{O}(K^2)$ time.

For enumerating all spanning trees of the equivalent graph, the algorithm proposed by Shioura and Tamura [1995] is used, which requires $\mathcal{O}(K + |G| + M) = \mathcal{O}(K^2 + M)$ computation time. This was proven to be optimal in time complexity. Shioura and Akihisa's algorithm starts from a spanning tree formed by depth-first search, then replaces one edge at a time using cycle structures in the graph, traversing the space of all spanning trees of the graph. Hence, computing R_{aMST} takes $\mathcal{O}(K^2 + M)$ time.

C Proofs for Lemmas and Theorems in Permutation Distributions

C.1 Proof of Lemma 1

Proof. Define

$$R_A = \sum_{u=1}^K \frac{1}{m_u} \sum_{i,j \in \mathcal{C}_u} I_{g_i \neq g_j},$$

and

$$R_B = \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in \mathcal{C}_u, j \in \mathcal{C}_v} I_{g_i \neq g_j}.$$

We have

$$\begin{aligned} \mathbf{E}_{\mathbf{P}}[R_{C_0}] &= \mathbf{E}_{\mathbf{P}}[R_A] + \mathbf{E}_{\mathbf{P}}[R_B] \\ &= \sum_{u=1}^K \frac{1}{m_u} \sum_{i,j \in \mathcal{C}_u} \mathbf{P}_{\mathbf{P}}(g_i \neq g_j) + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in \mathcal{C}_u, j \in \mathcal{C}_v} \mathbf{P}_{\mathbf{P}}(g_i \neq g_j). \end{aligned}$$

Since $\mathbf{P}_P(g_i \neq g_j) = \begin{cases} 0 & \text{if } i = j \\ \frac{2n_a n_b}{N(N-1)} & \text{if } i \neq j \end{cases}$, thus

$$\begin{aligned} \mathbf{E}_P[R_{C_0}] &= \sum_{u=1}^K \frac{1}{m_u} m_u(m_u - 1) \frac{2n_a n_b}{N(N-1)} + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} m_u m_v \frac{2n_a n_b}{N(N-1)} \\ &= (N - K + |C_0|) \frac{2n_a n_b}{N(N-1)}. \end{aligned}$$

Now, to compute the second moment, first note that

$$\mathbf{E}_P[R_{C_0}^2] = \mathbf{E}_P[R_A^2] + \mathbf{E}_P[R_B^2] + 2\mathbf{E}_P[R_A R_B].$$

Expanding the right-hand-side in above,

$$\begin{aligned} \mathbf{E}_P[R_A^2] &= \sum_{u,v=1}^k \frac{1}{m_u m_v} \sum_{i,j \in \mathcal{C}_u, k,l \in \mathcal{C}_v} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l), \\ \mathbf{E}_P[R_B^2] &= \sum_{(u,v) \in C_0} \frac{1}{m_u^2 m_v^2} \sum_{i,k \in \mathcal{C}_u, j,l \in \mathcal{C}_v} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l) \\ &\quad + 2 \sum_{\{(u,v),(w,y)\} \subset C_0} \frac{1}{m_u m_v m_w m_y} \sum_{i \in \mathcal{C}_u, j \in \mathcal{C}_v, k \in \mathcal{C}_w, l \in \mathcal{C}_y} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l), \\ \mathbf{E}_P[R_A R_B] &= \sum_{u=1}^K \sum_{(v,w) \in C_0} \frac{1}{m_u m_v m_w} \sum_{i,j \in \mathcal{C}_u, k \in \mathcal{C}_v, l \in \mathcal{C}_w} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l). \end{aligned}$$

Since

$$\mathbf{P}_P(g_i \neq g_j, g_k \neq g_l) = \begin{cases} 0 & \text{if } i = j \text{ and/or } k = l \\ \frac{2n_a n_b}{N(N-1)} = 2p_1 & \text{if } \begin{cases} i = k, j = l, i \neq j \\ i = l, j = k, i \neq j \end{cases} \\ \frac{n_a n_b}{N(N-1)} = p_1 & \text{if } \begin{cases} i = k, j \neq i, l \\ i = l, j \neq i, k \\ j = k, i \neq j, l \\ j = l, i \neq j, k \end{cases} \\ \frac{4n_a(n_a-1)n_b(n_b-1)}{N(N-1)(N-2)(N-3)} = p_2 & \text{if } i, j, k, l \text{ are all different,} \end{cases}$$

we have

$$\mathbf{E}_P[R_A^2] = \sum_{u=1}^K \frac{1}{m_u^2} \sum_{i,j,k,l \in \mathcal{C}_u} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l) + \sum_{u=1}^k \sum_{v \neq u} \frac{1}{m_u m_v} \sum_{i,j \in \mathcal{C}_u, k,l \in \mathcal{C}_v} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l)$$

$$\begin{aligned}
&= \sum_{u=1}^K \frac{1}{m_u^2} [2m_u(m_u - 1)(2p_1) + 4m_u(m_u - 1)(m_u - 2)p_1 + m_u(m_u - 1)(m_u - 2)(m_u - 3)p_2] \\
&\quad + \sum_{u=1}^k \sum_{v \neq u} \frac{1}{m_u m_v} m_u(m_u - 1)m_v(m_v - 1)p_2 \\
&= 4 \left(N - 2K + \sum_{u=1}^K \frac{1}{m_u} \right) p_1 + (N - K - 4)(N - K)p_2 + 6 \left(K - \sum_{u=1}^K \frac{1}{m_u} \right) p_2,
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}_P[R_B^2] &= \sum_{(u,v) \in C_0} \frac{1}{m_u^2 m_v^2} \sum_{i,k \in \mathcal{C}_u, j,l \in \mathcal{C}_v} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l) \\
&\quad + \sum_{(u,v), (u,w) \in C_0, v \neq w} \frac{1}{m_u^2 m_v m_w} \sum_{i,k \in \mathcal{C}_u, j \in \mathcal{C}_v, l \in \mathcal{C}_w} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l) \\
&\quad + \sum_{\substack{(u,v), (w,y) \in C_0 \\ u, v, w, y \text{ all different}}} \frac{1}{m_u m_v m_w m_y} \sum_{\substack{i \in \mathcal{C}_u, j \in \mathcal{C}_v \\ k \in \mathcal{C}_w, l \in \mathcal{C}_y}} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l) \\
&= \sum_{(u,v) \in C_0} \frac{1}{m_u^2 m_v^2} [m_u m_v (2p_1) + m_u m_v (m_u + m_v - 2)p_1 + m_u(m_u - 1)m_v(m_v - 1)p_2] \\
&\quad + \sum_{(u,v), (u,w) \in C_0, v \neq w} \frac{1}{m_u^2 m_v m_w} [m_u m_v m_w p_1 + m_u(m_u - 1)m_v m_w p_2] \\
&\quad + \sum_{\substack{(u,v), (w,y) \in C_0 \\ u, v, w, y \text{ all different}}} \frac{1}{m_u m_v m_w m_y} m_u m_v m_w m_y p_2 \\
&= \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} [(m_u + m_v)p_1 + (m_u - 1)(m_v - 1)p_2] \\
&\quad + \sum_{(u,v), (u,w) \in C_0, v \neq w} \frac{1}{m_u} [p_1 + (m_u - 1)p_2] \\
&\quad + 2|\{(u,v), (w,y)\} \subset C_0 : u, v, w, y \text{ all different}|p_2 \\
&= \sum_{u=1}^K \frac{|\mathcal{E}_u^{C_0}|^2}{m_u} (p_1 - p_2) + |C_0|^2 p_2 + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} p_2,
\end{aligned}$$

$$\mathbf{E}_P[R_A R_B] = \sum_{u=1}^K \sum_{(u,v) \in \mathcal{E}_u^{C_0}} \frac{1}{m_u^2 m_v} \sum_{i,j,k \in \mathcal{C}_u, l \in \mathcal{C}_w} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l)$$

$$\begin{aligned}
& + \sum_{u=1}^K \sum_{(v,w) \in C_0 \setminus \mathcal{E}_u^{C_0}} \frac{1}{m_u m_v m_w} \sum_{i,j \in \mathcal{C}_u} \sum_{k \in \mathcal{C}_v, l \in \mathcal{C}_w} \mathbf{P}_{\mathbf{P}}(g_i \neq g_j, g_k \neq g_l) \\
& = \sum_{u=1}^K \sum_{(u,v) \in \mathcal{E}_u^{C_0}} \frac{1}{m_u^2 m_v} [2m_u(m_u - 1)m_v p_1 + m_u(m_u - 1)(m_u - 2)m_v p_2] \\
& \quad + \sum_{u=1}^K \sum_{(v,w) \in C_0 \setminus \mathcal{E}_u^{C_0}} \frac{1}{m_u m_v m_w} m_u(m_u - 1)m_v m_w p_2 \\
& = |C_0|(N - K)p_2 + 2(p_1 - p_2) \left(2|C_0| - \frac{|\mathcal{E}_u^{C_0}|}{m_u} \right).
\end{aligned}$$

$\mathbf{Var}_{\mathbf{P}}[R_{C_0}]$ follows by combining the above in computing $\mathbf{E}_{\mathbf{P}}[R_{C_0}^2]$, and then subtracting $\mathbf{E}_{\mathbf{P}}^2[R_{C_0}]$. \square

C.2 Proof of Theorem 3

To prove Theorem 3, we first prove a simpler result: Asymptotic normality of the statistic under the bootstrap null, defined as the distribution obtained by sampling the group labels from the observed vector of group labels *with replacement*. Let $\mathbf{P}_{\mathbf{B}}$, $\mathbf{E}_{\mathbf{B}}$ and $\mathbf{Var}_{\mathbf{B}}$ denote respectively the probability, expectation and variance under the bootstrap null.

Lemma 3. *Assuming condition 1, under the bootstrap null distribution, the standardized statistic*

$$\frac{R_{C_0} - \mathbf{E}_{\mathbf{B}}[R_{C_0}]}{\sqrt{\mathbf{Var}_{\mathbf{B}}[R_{C_0}]}}$$

converges in distribution to $N(0, 1)$ as $K \rightarrow \infty$, where $\mathbf{E}_{\mathbf{B}}[R_{C_0}]$ and $\mathbf{Var}_{\mathbf{B}}[R_{C_0}]$ are given below.

$$\mathbf{E}_{\mathbf{B}}[R_{C_0}] = (N - K + |C_0|)2p_3, \quad (27)$$

$$\begin{aligned}
\mathbf{Var}_{\mathbf{B}}[R_{C_0}] & = 4(p_3 - p_4) \left(N - K + 2|C_0| + \sum_{u=1}^K \frac{|\mathcal{E}_u^{C_0}|^2}{4m_u} - \sum_{u=1}^K \frac{|\mathcal{E}_u^{C_0}|}{m_u} \right) \\
& \quad + (6p_4 - 4p_3) \left(K - \sum_{u=1}^K \frac{1}{m_u} \right) + p_4 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v},
\end{aligned} \quad (28)$$

where

$$p_3 = \frac{n_a n_b}{N^2}, \quad p_4 = \frac{4n_a^2 n_b^2}{N^4} = 4p_3^2. \quad (29)$$

The proof of Lemma 3 relies on Stein's method. Consider sums of the form $W = \sum_{i \in \mathcal{J}} \xi_i$, where \mathcal{J} is an index set and ξ are random variables with $E[\xi_i] = 0$, and $E[W^2] = 1$. The following assumption restricts the dependence between $\{\xi_i : i \in \mathcal{J}\}$.

Assumption 2. [Chen and Shao, 2005, p. 17] For each $i \in \mathcal{J}$ there exists $S_i \subset T_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{S_i^c}$ and ξ_{S_i} is independent of $\xi_{T_i^c}$.

We will use the following existing theorem.

Theorem 5. [Chen and Shao, 2005, Theorem 3.4] Under Assumption 2, we have

$$\sup_{h \in \text{Lip}(1)} |\mathbf{E}h(W) - \mathbf{E}h(Z)| \leq \delta$$

where $\text{Lip}(1) = \{h : \mathbb{R} \rightarrow \mathbb{R}\}$, Z has $\mathcal{N}(0, 1)$ distribution and

$$\delta = 2 \sum_{i \in \mathcal{J}} (\mathbf{E}|\xi_i \eta_i \theta_i| + |\mathbf{E}(\xi_i \eta_i)| \mathbf{E}|\theta_i|) + \sum_{i \in \mathcal{J}} \mathbf{E}|\xi_i \eta_i^2|$$

with $\eta_i = \sum_{j \in S_i} \xi_j$ and $\theta_i = \sum_{j \in T_i} \xi_j$, where S_i and T_i are defined in Assumption 2.

Proof of Lemma 3. The mean and variance of R_{C_0} under the bootstrap null, (27) and (28), can be obtained following similar steps as the proof of Lemma 1, noting that, under the bootstrap null,

$$\mathbf{P}_B(g_i \neq g_j) = \begin{cases} 0 & \text{if } i = j \\ \frac{2n_a n_b}{N^2} = 2p_3 & \text{if } i \neq j \end{cases},$$

and

$$\mathbf{P}_B(g_i \neq g_j, g_k \neq g_l) = \begin{cases} 0 & \text{if } i = j \text{ and/or } k = l \\ \frac{2n_a n_b}{N^2} = 2p_3 & \text{if } \begin{cases} i = k, j = l, i \neq j \\ i = l, j = k, i \neq j \end{cases} \\ \frac{n_a n_b}{N^2} = p_3 & \text{if } \begin{cases} i = k, j \neq i, l \\ i = l, j \neq i, k \\ j = k, i \neq j, l \\ j = l, i \neq j, k \end{cases} \\ \frac{4n_a^2 n_b^2}{N^4} = p_4 & \text{if } i, j, k, l \text{ are all different} \end{cases}.$$

To prove asymptotic normality, we first define more notations. For any node u of C_0 , let

$$R_u = \frac{2n_{au}n_{bu}}{m_u}, \quad d_u = \mathbf{E}_B[R_u] = 2(m_u - 1)p_3,$$

where p_3 is defined in (29). Similarly, for any edge (u, v) of C_0 , let

$$R_{uv} = \frac{n_{au}n_{bv} + n_{av}n_{bu}}{m_um_v}, \quad d_{uv} = \mathbf{E}_B[R_{uv}] = 2p_3.$$

Let $\sigma_B^2 = \mathbf{Var}_B[R_{C_0}]$, ξ_u, ξ_{uv} be the standardized mixing potentials,

$$\xi_u = \frac{R_u - d_u}{\sigma_B}, \quad (30)$$

$$\xi_{uv} = \frac{R_{uv} - d_{uv}}{\sigma_B}. \quad (31)$$

Finally, we define the index sets for ξ_u and ξ_{uv} :

$$\mathcal{J}_1 = \{1, \dots, K\},$$

$$\mathcal{J}_2 = \{uv : u < v \text{ such that } (u, v) \in C_0\},$$

and let $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2$. Since $R_{C_0} = \sum_{u=1}^K R_u + \sum_{(u,v) \in C_0} R_{uv}$, the standardized statistic is

$$W := \sum_{i \in \mathcal{J}} \xi_i = \sum_{u \in \mathcal{J}_1} \frac{R_u - d_u}{\sigma_B} + \sum_{uv \in \mathcal{J}_2} \frac{R_{uv} - d_{uv}}{\sigma_B} = \frac{R_{C_0} - \mathbf{E}_B[R_{C_0}]}{\sigma_B}.$$

Our notation follows those of Theorem 5 and Assumption 2. For $u \in \mathcal{J}_1$, let

$$\begin{aligned} S_u &= \{u\} \cup \{uv, vu : (u, v) \in C_0\}, \\ T_u &= S_u \cup \{v, vw, wv : (u, v), (v, w) \in C_0\}. \end{aligned}$$

For $uv \in \mathcal{J}_2$, let

$$\begin{aligned} S_{uv} &= \{uv, u, v\} \cup \{uw, wu : (u, w) \in C_0\} \cup \{vw, wv : (v, w) \in C_0\}, \\ T_{uv} &= S_{uv} \cup \{w, wy, yw : (u, w), (w, y) \in C_0\} \cup \{w, wy, yw : (v, w), (w, y) \in C_0\}. \end{aligned}$$

S_u, T_u, S_{uv}, T_{uv} defined in this way satisfy Assumption 2.

Since $R_u \in [0, \frac{m_u}{2}]$, $p_3 \in [0, \frac{1}{4}]$, and $R_{uv} \in [0, 1]$, we have $d_u \in [0, \frac{m_u-1}{2}]$, $d_{uv} \in [0, \frac{1}{2}]$, and therefore $|\xi_u| \leq \frac{m_u}{2\sigma_B}$, $|\xi_{uv}| \leq \frac{1}{\sigma_B}$. Hence,

$$\begin{aligned} \sum_{j \in S_u} |\xi_j| &\leq \frac{1}{\sigma_B}(m_u + |\mathcal{E}_u^{C_0}|), \quad u \in \mathcal{J}_1, \\ \sum_{j \in T_u} |\xi_j| &\leq \frac{1}{\sigma_B}(m_u + \sum_{v \in \mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|), \quad u \in \mathcal{J}_1, \\ \sum_{j \in S_{uv}} |\xi_j| &\leq \frac{1}{\sigma_B}(m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|), \quad uv \in \mathcal{J}_2, \\ \sum_{j \in T_{uv}} |\xi_j| &\leq \frac{1}{\sigma_B}(m_u + m_v + \sum_{w \in \mathcal{V}_u \cup \mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|), \quad uv \in \mathcal{J}_2. \end{aligned}$$

As in Theorem 5, let $\eta_i = \sum_{j \in S_i} \xi_j$ and $\theta_i = \sum_{j \in T_i} \xi_j$. Then

$$\begin{aligned} \mathbf{E}_B |\xi_i \eta_i \theta_i| &= \mathbf{E}_B |\xi_i| \sum_{j \in S_i} \xi_j \sum_{k \in T_i} \xi_k \leq \mathbf{E}_B |\xi_i| \sum_{j \in S_i} |\xi_j| \sum_{k \in T_i} |\xi_k|, \\ |\mathbf{E}_B(\xi_i \eta_i)| \mathbf{E}_B |\theta_i| &\leq \mathbf{E}_B |\xi_i| \sum_{j \in S_i} \xi_j |\mathbf{E}_B| \sum_{j \in T_i} \xi_j \leq \mathbf{E}_B |\xi_i| \sum_{j \in S_i} |\xi_j| \mathbf{E}_B \sum_{j \in T_i} |\xi_j|, \\ \mathbf{E}_B |\xi_i \eta_i^2| &= \mathbf{E}_B |\xi_i| \sum_{j \in S_i} \sum_{k \in S_i} \xi_j \xi_k \leq \mathbf{E}_B |\xi_i| \sum_{j \in S_i} |\xi_j| \sum_{k \in S_i} |\xi_k|. \end{aligned}$$

Thus, for $i = u \in \mathcal{J}_1$, the terms $\mathbf{E}_B |\xi_i \eta_i \theta_i|$, $|\mathbf{E}_B(\xi_i \eta_i)| \mathbf{E}_B |\theta_i|$, and $\mathbf{E}_B |\xi_i \eta_i^2|$ are all bounded by

$$\frac{1}{\sigma_B^3} m_u (m_u + |\mathcal{E}_u^{C_0}|) (m_u + \sum_{v \in \mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|),$$

and for $i = uv \in \mathcal{J}_2$, the terms $\mathbf{E}_B |\xi_i \eta_i \theta_i|$, $|\mathbf{E}_B(\xi_i \eta_i)| \mathbf{E}_B |\theta_i|$, and $\mathbf{E}_B |\xi_i \eta_i^2|$ are all bounded by

$$\frac{1}{\sigma_B^3} (m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|) (m_u + m_v + \sum_{w \in \mathcal{V}_u \cup \mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|).$$

Hence,

$$\begin{aligned} \delta &\leq \frac{5}{\sigma_B^3} \left(\sum_{u=1}^K m_u (m_u + |\mathcal{E}_u^{C_0}|) (m_u + \sum_{v \in \mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|) \right. \\ &\quad \left. + \sum_{(u,v) \in C_0} (m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|) (m_u + m_v + \sum_{w \in \mathcal{V}_u \cup \mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|) \right). \end{aligned}$$

Since $\sigma_{\mathbf{B}}$ is of order \sqrt{K} or higher, under condition 1, $\delta \rightarrow 0$ as $K \rightarrow \infty$.

□

Proof of Theorem 3. To show the asymptotic normality of the standardized statistic under the permutation null, we only need to show that (R_{C_0}, n_a^B) converges to a bivariate Gaussian distribution under the bootstrap null, where n_a^B is the number of observations that belong to group a in the bootstrap sample. Then asymptotic normality of R_{C_0} under the permutation null follows from the fact that its distribution is equal to the conditional distribution of R_{C_0} given $n_a^B = n_a$. The standardized bivariate vector is

$$\left(\frac{R_{C_0} - \mathbf{E}_{\mathbf{B}}[R_{C_0}]}{\sqrt{\mathbf{Var}_{\mathbf{B}}[R_{C_0}]}} , \frac{n_a^B - Np_a}{\sigma_0} \right)$$

with $p_a = n_a/N, \sigma_0^2 = Np_a(1 - p_a)$. By the Cramér-Wold device, we only need to show that

$$a_1 \frac{R_{C_0} - \mathbf{E}_{\mathbf{B}}[R_{C_0}]}{\sqrt{\mathbf{Var}_{\mathbf{B}}[R_{C_0}]}} + a_2 \frac{n_a^B - Np_a}{\sigma_0}$$

is asymptotic Gaussian under the bootstrap null for all $a_1, a_2 \in \mathbb{R}, a_1 a_2 \neq 0$.

Let $\xi_i, i \in \mathcal{J}$ be defined in the same way as in the proof of Lemma 3. Let $\mathcal{J}_3 = \{|\mathcal{J}| + 1, \dots, |\mathcal{J}| + K\}$. For $i \in \mathcal{J}_3$, let

$$\xi_i = \frac{n_{ai'} - p_a m_{i'}}{\sigma_0}, \quad i' = i - |\mathcal{J}|.$$

We use Theorem 5 to show the asymptotic Gaussianity of $\sum_{i \in \mathcal{J}} a_1 \xi_i + \sum_{i \in \mathcal{J}_3} a_2 \xi_i$. We need to redefine the neighborhood sets to satisfy Assumption 2.

For $u \in \mathcal{J}_1$,

$$\begin{aligned} S_u &= \{u, u + |\mathcal{J}|\} \cup \{uv, vu : (u, v) \in C_0\}, \\ T_u &= S_u \cup \{v, v + |\mathcal{J}|, vw, wv : (u, v), (v, w) \in C_0\}. \end{aligned}$$

For $uv \in \mathcal{J}_2$,

$$\begin{aligned} S_{uv} &= \{uv, u, v, u + |\mathcal{J}|, v + |\mathcal{J}|\} \cup \{uw, wu : (u, w) \in C_0\} \\ &\quad \cup \{vw, wv : (v, w) \in C_0\}, \\ T_{uv} &= S_{uv} \cup \{w, w + |\mathcal{J}|, wy, yw : (u, w), (w, y) \in C_0\} \\ &\quad \cup \{w, w + |\mathcal{J}|, wy, yw : (v, w), (w, y) \in C_0\}. \end{aligned}$$

And for $u \in \mathcal{J}_3$,

$$S_u = \{u, u'\} \cup \{u'v, vu' : (u', v) \in C_0\}, \quad u' = u - |\mathcal{J}|,$$

$$T_u = S_u \cup \{v, v + |\mathcal{J}|, vw, wv : (u', v), (v, w) \in C_0\}.$$

From the proof of Lemma 3, we have

$$|\xi_u| \leq \frac{m_u}{2\sigma_B}, \quad \forall u \in \mathcal{J}_1; \quad |\xi_{uv}| \leq \frac{1}{\sigma_B}, \quad \forall uv \in \mathcal{J}_2.$$

For $u \in \mathcal{J}_3$,

$$|\xi_u| \leq \frac{m_{u'}}{\sigma_0}, \quad u' = u - |\mathcal{J}|.$$

Let $\sigma = \min(\sigma_B, \sigma_0)$, then

$$\sum_{j \in S_u} |\xi_j| \leq \frac{1}{\sigma}(2m_u + |\mathcal{E}_u^{C_0}|), \quad u \in \mathcal{J}_1 \cup \mathcal{J}_3,$$

$$\sum_{j \in T_u} |\xi_j| \leq \frac{1}{\sigma}(2m_u + 2 \sum_{v \in \mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|), \quad u \in \mathcal{J}_1 \cup \mathcal{J}_3,$$

$$\sum_{j \in S_{uv}} |\xi_j| \leq \frac{1}{\sigma}(2m_u + 2m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|), \quad uv \in \mathcal{J}_2,$$

$$\sum_{j \in T_{uv}} |\xi_j| \leq \frac{1}{\sigma}(2m_u + 2m_v + 2 \sum_{w \in \mathcal{V}_u \cup \mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|), \quad uv \in \mathcal{J}_2.$$

Thus, for $i = u \in \mathcal{J}_1 \cup \mathcal{J}_3$, the terms $\mathbf{E}_B|\xi_i \eta_i \theta_i|$, $|\mathbf{E}_B(\xi_i \eta_i)|\mathbf{E}_B|\theta_i|$, and $\mathbf{E}_B|\xi_i \eta_i^2|$ are all bounded by

$$\frac{1}{\sigma^3} m_u (2m_u + |\mathcal{E}_u^{C_0}|) (2m_u + 2 \sum_{v \in \mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|),$$

and for $i = uv \in \mathcal{J}_2$, terms $\mathbf{E}_B|\xi_i \eta_i \theta_i|$, $|\mathbf{E}_B(\xi_i \eta_i)|\mathbf{E}_B|\theta_i|$, and $\mathbf{E}_B|\xi_i \eta_i^2|$ are all bounded by

$$\frac{1}{\sigma^3} (2m_u + 2m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|) (2m_u + 2m_v + 2 \sum_{w \in \mathcal{V}_u \cup \mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|).$$

Define $W_{a_1, a_2} = \sum_{i \in \mathcal{J}} a_1 \xi_i + \sum_{i \in \mathcal{J}_3} a_2 \xi_i$. The value of δ in Theorem 5 has the form

$$\delta = \frac{1}{\sqrt{\mathbf{E}_B[W_{a_1, a_2}^2]}} \left(2 \sum_{i \in \mathcal{J}} (\mathbf{E}_B|a_1 \xi_i \eta_i \theta_i| + |\mathbf{E}_B(a_1 \xi_i \eta_i)|\mathbf{E}_B|\theta_i|) + \sum_{i \in \mathcal{J}} \mathbf{E}_B|a_1 \xi_i \eta_i^2| \right. \\ \left. + 2 \sum_{i \in \mathcal{J}_3} (\mathbf{E}_B|a_2 \xi_i \eta_i \theta_i| + |\mathbf{E}_B(a_2 \xi_i \eta_i)|\mathbf{E}_B|\theta_i|) + \sum_{i \in \mathcal{J}_3} \mathbf{E}_B|a_2 \xi_i \eta_i^2| \right),$$

where $\eta_i = \sum_{j \in S_i} \xi_j (a_1 I_{j \in \mathcal{J}} + a_2 I_{j \in \mathcal{J}_3})$, and $\theta_i = \sum_{j \in T_i} \xi_j (a_1 I_{j \in \mathcal{J}} + a_2 I_{j \in \mathcal{J}_3})$.

Let $a = \max(|a_1|, |a_2|)$, we have

$$\begin{aligned}
\mathbf{E}_{\mathbf{B}}|a_1 \xi_i \eta_i \theta_i|, \mathbf{E}_{\mathbf{B}}|a_2 \xi_i \eta_i \theta_i| &\leq a^3 \mathbf{E}_{\mathbf{B}}|\xi_i| \sum_{j \in S_i} \xi_j \sum_{k \in T_i} \xi_k| \\
&\leq a^3 \mathbf{E}_{\mathbf{B}}|\xi_i| \sum_{j \in S_i} |\xi_j| \sum_{k \in T_i} |\xi_k|, \\
|\mathbf{E}_{\mathbf{B}}(a_1 \xi_i \eta_i)| \mathbf{E}_{\mathbf{B}}|\theta_i|, |\mathbf{E}_{\mathbf{B}}(a_2 \xi_i \eta_i)| \mathbf{E}_{\mathbf{B}}|\theta_i| &\leq a^3 \mathbf{E}_{\mathbf{B}}|\xi_i| \sum_{j \in S_i} \xi_j |\mathbf{E}_{\mathbf{B}}| \sum_{j \in T_i} \xi_j| \\
&\leq a^3 \mathbf{E}_{\mathbf{B}}|\xi_i| \sum_{j \in S_i} |\xi_j| \mathbf{E}_{\mathbf{B}} \sum_{j \in T_i} |\xi_j|, \\
\mathbf{E}_{\mathbf{B}}|a_1 \xi_i \eta_i^2|, \mathbf{E}_{\mathbf{B}}|a_2 \xi_i \eta_i^2| &\leq a^3 \mathbf{E}_{\mathbf{B}}|\xi_i| \sum_{j \in S_i} \sum_{k \in S_i} \xi_j \xi_k| \\
&\leq a^3 \mathbf{E}_{\mathbf{B}}|\xi_i| \sum_{j \in S_i} |\xi_j| \sum_{k \in S_i} |\xi_k|.
\end{aligned}$$

Thus,

$$\begin{aligned}
\delta &\leq \frac{40a^3}{\sigma^3 \sqrt{\mathbf{E}_{\mathbf{B}}[W_{a_1, a_2}^2]}} \left(\sum_{u=1}^K m_u (m_u + |\mathcal{E}_u^{C_0}|) (m_u + \sum_{v \in \mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|) \right. \\
&\quad \left. + \sum_{(u,v) \in C_0} (m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|) (m_u + m_v + \sum_{w \in \mathcal{V}_u \cup \mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|) \right).
\end{aligned}$$

Since $\sigma_{\mathbf{B}}^2$ is at least of order K and σ_0^2 is of order N , σ^2 is at least of order K by Condition 2. If $\mathbf{E}_{\mathbf{B}}[W_{a_1, a_2}^2]$ is uniformly strictly bounded from 0 for any $a_1 a_2 \neq 0$, then under Condition 1, $\delta \rightarrow 0$ as $K \rightarrow \infty$.

We next show that under Condition 2, $\mathbf{E}_{\mathbf{B}}[W_{a_1, a_2}^2]$ is uniformly strictly bounded from 0 for any $a_1 a_2 \neq 0$.

Let $W_1 = \sum_{i \in \mathcal{J}} \xi_i$, $W_2 = \sum_{i \in \mathcal{J}_3} \xi_i$, then

$$\begin{aligned}
\mathbf{E}_{\mathbf{B}}[W_{a_1, a_2}^2] &= a_1^2 \mathbf{E}_{\mathbf{B}} W_1^2 + a_2^2 \mathbf{E}_{\mathbf{B}} W_2^2 + 2a_1 a_2 \mathbf{E}_{\mathbf{B}}[W_1 W_2] \\
&= a_1^2 + a_2^2 + 2a_1 a_2 \mathbf{E}_{\mathbf{B}}[W_1 W_2]
\end{aligned}$$

Thus, we only need to show that the absolute correlation between W_1 and W_2 is uniformly strictly bounded from 1. Notice that, in the theorem, we require n_a/N to be bounded from 0 and 1, so p_a and p_b are both bounded from 0 and 1.

Correlation between R_{C_0} and n_a^B : Observe that

$$\begin{aligned} R_{C_0} n_a^B &= \left[\sum_{u=1}^K \frac{1}{m_u} \sum_{i,j \in \mathcal{C}_u} I_{g_i \neq g_j} + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in \mathcal{C}_u, j \in \mathcal{C}_v} I_{g_i \neq g_j} \right] \sum_{x=1}^N I_{g_x=a} \\ &= \sum_{u=1}^K \frac{1}{m_u} \sum_{i,j \in \mathcal{C}_u} \left(I_{g_i \neq g_j} \sum_{x=1}^N I_{g_x=a} \right) + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in \mathcal{C}_u, j \in \mathcal{C}_v} \left(I_{g_i \neq g_j} \sum_{x=1}^N I_{g_x=a} \right). \end{aligned}$$

For any $i \neq j$,

$$\begin{aligned} \mathbf{E}_B \left[I_{g_i \neq g_j} \sum_{x=1}^N I_{g_x=a} \right] &= \mathbf{E}_B \left[I_{g_i \neq g_j, g_i=a} + I_{g_i \neq g_j, g_j=a} + \sum_{x \neq i,j} I_{g_i \neq g_j, g_x=a} \right] \\ &= \mathbf{P}_B(g_i = a, g_j = b) + \mathbf{P}_B(g_i = b, g_j = a) + \sum_{x \neq i,j} \mathbf{P}_B(g_i \neq g_j, g_x = a) \\ &= p_a p_b + p_a p_b + 2p_a p_b (N-2) = 2p_a p_b (N p_a + 1 - 2p_a). \end{aligned}$$

Hence

$$\mathbf{E}_B[R_{C_0} n_a^B] = (N - K + |C_0|) 2p_a p_b (N p_a + 1 - 2p_a).$$

Since $\mathbf{E}_B[R_{C_0}] = (N - K + |C_0|) 2p_a p_b$ and $\mathbf{E}_B[n_a^B] = N p_a$, we have

$$\mathbf{Cov}_B(R_{C_0}, n_a^B) = (N - K + |\mathcal{E}_{C_0}|) 2p_a p_b (1 - 2p_a). \quad (32)$$

If $p_a = 1/2$, then $\mathbf{Cov}_B(R_{C_0}, n_a^B) = 0$. Since $\mathbf{Var}_B[R_{C_0}]$ and $\mathbf{Var}_B[n_a^B] = N p_a p_b$ are positive, $\mathbf{Cor}_B(R_{C_0}, n_a^B) = 0$, clearly bounded from 1. We consider $p_a \neq 1/2$ in the following.

$$\begin{aligned} \mathbf{Var}_B[R_{C_0}] &= 4p_a p_b (1 - 4p_a p_b) \left(N - K + 2|C_0| + \sum_{u=1}^K \frac{|\mathcal{E}_u^{C_0}|^2}{4m_u} - \sum_{u=1}^K \frac{|\mathcal{E}_u^{C_0}|}{m_u} \right) \\ &\quad + 4p_a p_b (6p_a p_b - 1) \left(K - \sum_{u=1}^K \frac{1}{m_u} \right) + 4p_a^2 p_b^2 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \\ &= 4p_a p_b (1 - 4p_a p_b) \left(N - 2K + 2|C_0| + \sum_{u=1}^K \frac{(|\mathcal{E}_u^{C_0}|/2 - 1)^2}{m_u} \right) \\ &\quad + 8p_a^2 p_b^2 \left(K - \sum_{u=1}^K \frac{1}{m_u} \right) + 4p_a^2 p_b^2 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}. \end{aligned}$$

Since

$$\begin{aligned} N \sum_{u=1}^K \frac{(|\mathcal{E}_u^{C_0}|/2 - 1)^2}{m_u} &= \sum_{u=1}^K m_u \sum_{u=1}^K \frac{(|\mathcal{E}_u^{C_0}|/2 - 1)^2}{m_u} \geq \left(\sum_{u=1}^K \sqrt{m_u \frac{(|\mathcal{E}_u^{C_0}|/2 - 1)^2}{m_u}} \right)^2 \\ &= \left(\sum_{u=1}^K \left| |\mathcal{E}_u^{C_0}|/2 - 1 \right| \right)^2 \geq \left(\sum_{u=1}^K (|\mathcal{E}_u^{C_0}|/2 - 1) \right)^2 = (|C_0| - K)^2, \end{aligned}$$

we have

$$\mathbf{Var}_B[R_{C_0}] \mathbf{Var}_B[n_a^B] \geq 4p_a^2 p_b^2 (1 - 4p_a p_b) [N - K + |C_0|]^2 + 4p_a^3 p_b^3 N \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}.$$

Hence,

$$|\mathbf{Cor}_B(R_{C_0}, n_a^B)| \leq \frac{1}{\sqrt{1 + \frac{p_a p_b N \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}}{(1 - 4p_a p_b) [N - K + |C_0|]^2}}}.$$

When $N, |C_0|, \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sim \mathcal{O}(K)$, $|\mathbf{Cor}_B(R_{C_0}, n_a^B)|$ is bounded by a value smaller than 1.

□

C.3 Proof of Lemma 2

Let \bar{G} be the uMST on subjects, and $\mathcal{E}_i^{\bar{G}} = \{(i, j) : (i, j) \in \bar{G}\}$. Then $|\mathcal{E}_i^{\bar{G}}| = m_u + \sum_{\mathcal{V}_u} m_v - 1$, $|\bar{G}| = \sum_{u=1}^K m_u(m_u - 1)/2 + \sum_{(u,v) \in C_0} m_u m_v$. Since $\mathbf{E}_P[T_{C_0}] = |\bar{G}|/2p_1$, and the result follows.

Now, we compute the second moment.

$$\begin{aligned} \mathbf{E}_P[T_{C_0}^2] &= \sum_{(i,j),(k,l) \in \bar{G}} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l) \\ &= \sum_{(i,j) \in \bar{G}} \mathbf{P}_P(g_i \neq g_j) + \sum_{(i,j),(i,k) \in \bar{G}, j \neq k} \mathbf{P}_P(g_i \neq g_j, g_i \neq g_k) \\ &\quad + \sum_{\substack{(i,j),(k,l) \in \bar{G} \\ i,j,k,l \text{ all different}}} \mathbf{P}_P(g_i \neq g_j, g_k \neq g_l) \\ &= |\bar{G}|/2p_1 + \sum_{i=1}^N |\mathcal{E}_i^{\bar{G}}| (|\mathcal{E}_i^{\bar{G}}| - 1) p_1 + (|\bar{G}|^2 - |\bar{G}| - \sum_{i=1}^N |\mathcal{E}_i^{\bar{G}}| (|\mathcal{E}_i^{\bar{G}}| - 1)) p_2 \end{aligned}$$

$$\begin{aligned}
&= (p_1 - p_2) \sum_{u=1}^K m_u (m_u + \sum_{v \in \mathcal{V}_u} m_v - 1) (m_u + \sum_{v \in \mathcal{V}_u} m_v - 2) \\
&\quad + (p_1 - p_2/2) \left(\sum_{u=1}^K m_u (m_u - 1) + 2 \sum_{(u,v) \in C_0} m_u m_v \right) \\
&\quad + p_2 \left(\sum_{u=1}^K m_u (m_u - 1) + 2 \sum_{(u,v) \in C_0} m_u m_v \right)^2.
\end{aligned}$$

$\mathbf{Var}_{\mathbf{P}}[T_{C_0}]$ follows by $\mathbf{E}_{\mathbf{P}}[T_{C_0}^2] - \mathbf{E}_{\mathbf{P}}^2[T_{C_0}]$.

□

Acknowledgments

Hao Chen was supported by NSF DMS Grant 1043204 and an NIH Training Grant. Nancy R. Zhang was supported by NSF DMS Grant 0906394 and NIH Grant R01 HG006137-01. We thank one of the reviewers for bring to our attention Critchlow [1985]’s work.

References

- J.A. Anderson, K. Whaley, J. Williamson, and W.W. Buchanan. A statistical aid to the diagnosis of keratoconjunctivitis sicca. *QJM*, 41(2):175, 1972.
- E.C. Bush and B.T. Lahn. The evolution of word composition in metazoan promoter sequence. *PLoS Computational Biology*, 2:e150, 2006.
- H. Chen. *Graph-based tests*. PhD thesis, in progress, Stanford University, 2012.
- L.H.Y. Chen and Q.M. Shao. Stein’s method for normal approximation. *An introduction to Stein’s method*, 4:1–59, 2005.
- D.E. Critchlow. *Metric methods for analyzing partially ranked data*, volume 34. Springer, 1985.
- P. Diaconis. Group representations in probability and statistics. *Lecture Notes-Monograph Series*, 11, 1988.
- D. Eppstein. *Representing all minimum spanning trees with applications to counting and generation*. Citeseer, 1995.

- J.H. Friedman and L.C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979.
- S. Furihata, T. Ito, and N. Kamatani. Test of association between haplotypes and phenotypes in case-control studies: Examination of validity of the application of an algorithm for samples from cohort or clinical trials to case-control samples using simulated and real data. *Genetics*, 174(3):1505–1516, 2006.
- R.A. Lippert, H. Huang, and M.S. Waterman. Distributional regimes for the number of k-word matches between two random sequences. *Proceedings of the National Academy of Sciences*, 99(22):13980, 2002.
- C.L. Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.
- C.R. Mehta and N.R. Patel. A network algorithm for performing fisher’s exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association*, 78(382):427–434, 1983.
- D. Nettleton and T. Banerjee. Testing the equality of distributions of random vectors with categorical components. *Computational statistics & data analysis*, 37(2):195–208, 2001.
- S.C. Perry and R.G. Beiko. Distinguishing microbial genome fragments based on their composition: evolutionary and comparative genomic perspectives. *Genome Biology and Evolution*, 2:117–131, 2010.
- I. Rajan, S. Aravamuthan, and S.S. Mande. Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics*, 23: 2672–2677, 2007.
- P.R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005.
- Akiyoshi Shioura and Akihisa Tamura. Efficiently scanning all spanning trees of an undirected graph. *Journal of the Operations Research Society of Japan*, 38(3):331–344, 1995. ISSN 04534514. URL <http://ci.nii.ac.jp/naid/110001184429/en/>.

D.V. Zaykin, P.H. Westfall, S.S. Young, M.A. Karnoub, M.J. Wagner, and M.G. Ehm. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human heredity*, 53(2): 79–91, 2002.

Department of Statistics, Stanford University

E-mail: haochen@stanford.edu

Department of Statistics, The Wharton School, University of Pennsylvania

E-mail: nzh@wharton.upenn.edu